STATE SPACE ANALYSIS OF CONTROL SYSTEMS

PRENTICE-HALL ELECTRICAL ENGINEERING SERIES*
William L. Everitt, *Editor*

ANNER *Elements of Television Systems*
ARMINGTON & VOLZ *An Introduction to Electric Circuit Analysis*
BALABANIAN *Network Synthesis*
BARTON *Radar Systems Analysis*
BENEDICT *Introduction to Industrial Electronics*
BLACKWELL & KOTZEBUE *Semiconductor-Diode Parametric Amplifiers**
BOISVERT, ROBERT, & ROBICHAUD *Signal Flow Graphs and Applications*
CARLIN & GIORDANO *Network Theory: An Introduction to Reciprocal and Nonreciprocal Circuits*
CHANG, K. N. *Parametric and Tunnel Diodes*
CHANG, S. *Energy Conversion**
DAVIS & WEED *Industrial Electronic Engineering*
DEKKER *Electrical Engineering Materials*
DEL TORO *Principles of Electrical Engineering*
DE PIAN *Linear Active Network Theory*
DOWNING *Modulation Systems and Noise**
DUNN & BARKER *Electrical Measurements Manual*
EVANS *Experiments in Electronics*
EVANS *Introduction to Electronics*
FETT *Feedback Control Systems*
FICH *Transient Analysis in Electrical Engineering*
FICH & POTTER *Theory of A-C Circuits*
FLORES *Computer Logic: The Functional Design of Digital Computers*
FLORES *The Logic of Computer Arithmetic**
FOECKE *Introduction to Electrical Engineering Science*
GENTRY, ET AL. *Semiconductor Controlled Rectifiers: Principles and Applications of p-n-p-n Devices*
GOLDMAN *Information Theory*
GOLDMAN *Transformation Calculus and Electrical Transients*
GOLOMB, ET AL. *Digital Communications**
GRAY *Digital Computer Engineering**
HERSHBERGER *Principles of Communication Systems*
JORDAN *Electromagnetic Waves and Radiating Systems*
KUO *Analysis and Synthesis of Sampled-Data Control Systems**
KUO *Automatic Control Systems*
LECROISSETTE *Transistors*
LEGROS & MARTIN *Transform Calculus for Electrical Engineers*
LO, ET AL. *Transistor Electronics*
MALEY & EARLE *The Logic Design of Transistor Digital Computers**
MALEY & SKIKO *Modern Digital Computers*
MARCUS *Switching Circuits for Engineers*

* This title is also in the Prentice-Hall International Series in Electrical Engineering. Prentice-Hall, Inc.; Prentice-Hall International, Inc., United Kingdom and Eire; Prentice-Hall of Canada, Ltd., Canada.

MARTIN *Electronic Circuits*
MARTIN *Physical Basis for Electrical Engineering*
MARTIN *Ultrahigh Frequency Engineering*
MATSCH *Capacitors, Magnetic Circuits, and Transformers*
MOSKOWITZ & RACKER *Pulse Techniques*
NIXON *Principles of Automatic Controls*
NUSSBAUM *Electromagnetic and Quantum Properties of Materials*
NUSSBAUM *Electromagnetic Theory for Engineers and Scientists*
NUSSBAUM *Semiconductor Device Physics**
OGATA *State Space Analysis of Control Systems*
PARTRIDGE *Principles of Electronic Instruments*
PASKUSZ & BUSSELL *Linear Circuit Analysis*
PIERUSCHKA *Principles of Reliability**
POTTER & FICH *Theory of Networks and Lines*
PUMPHREY *Electrical Engineering, 2nd ed.*
PUMPHREY *Fundamentals of Electrical Engineering, 2nd ed.*
REED *Electric Networks Synthesis*
REED *Foundation for Electric Network Theory*
RIDEOUT *Active Networks*
ROBERTS & VANDERSLICE *Ultrahigh Vacuum and Its Applications*
ROBICHAUD, ET AL. *Signal Flow Graphs and Applications**
RUSSELL *Modulation and Coding in Information Systems*
RYDER, F. L. *Creative Engineering Analysis*
RYDER, J. D. *Electronic Engineering Principles, 3rd ed.*
RYDER, J. D. *Electronic Fundamentals and Applications, 3rd ed.*
RYDER, J. D. *Networks, Lines and Fields, 2nd ed.*
SANFORD *Physical Networks*
SHEDD *Fundamentals of Electromagnetic Waves*
SKRODER & HELM *Circuit Analysis by Laboratory Methods, 2nd ed.*
SOOHOO *Theory and Application of Ferrites*
STOUT *Basic Electrical Measurements, 2nd ed.*
THOMSON *Laplace Transformation, 2nd ed.*
VAN DER ZIEL *Noise*
VAN DER ZIEL *Solid State Physical Electronics*
VAN VALKENBURG *Network Analysis, 2nd ed.*
VON TERSCH & SWAGO *Recurrent Electrical Transients*
WALLMARK & JOHNSON (EDITORS) *Field-Effect Transistors: Physics Technology and Applications*
WARD *Introduction to Electrical Engineering, 2nd ed.*
WARFIELD *Introduction to Electronic Analog Computers*
WEED & DAVIS *Fundamentals of Electron Devices and Circuits*

INSTRUMENTATION AND CONTROLS SERIES

CHEN & HAAS
Elements of Control Systems Analysis
KUO
Discrete-Data Control Systems
KUO
Analysis and Synthesis of Sampled-Data Control Systems
OGATA
Modern Control Engineering
OGATA
State Space Analysis of Control Systems

PRENTICE-HALL INTERNATIONAL, INC., *London*
PRENTICE-HALL OF AUSTRALIA, PTY. LTD., *Sydney*
PRENTICE-HALL OF CANADA, LTD., *Toronto*
PRENTICE-HALL OF INDIA (PRIVATE) LTD., *New Delhi*
PRENTICE-HALL OF JAPAN, INC., *Tokyo*

STATE SPACE ANALYSIS OF CONTROL SYSTEMS

KATSUHIKO OGATA
University of Minnesota

PRENTICE-HALL, INC., *Englewood Cliffs, N.J.*

© 1967 by
PRENTICE-HALL, INC.
Englewood Cliffs, N.J.

All rights reserved. No part of this book may be reproduced in any form or by any means without permission in writing from the publisher.

Current Printing (Last Digit):
10 9 8 7

Library of Congress Catalog Card Number 67-14864
Printed in the United States of America
84453 C

Preface

This book presents an introductory account of state space analysis of control systems and furnishes a background necessary for studying modern control theory. It is primarily designed to be a textbook for a second course in the field of feedback control systems for seniors and first year graduate students majoring in the field of engineering. In addition to class-room use it will prove to be useful as a reference book for practicing control engineers who wish to study basic theory of state space analysis of control systems and to get some working knowledge of modern control theory. The prerequisite on the part of the reader is that he has had at least one course on feedback control systems at the junior-senior level.

Most of the material presented in this book has been offered in the author's courses at the senior and first year graduate level in the Department of Mechanical Engineering, University of Minnesota.

The first five chapters present essentially a background material and the last four chapters give main subjects of state space analysis of control systems. Specifically, Chapter 1 presents the introductory material. Chapter 2 reviews basic aspects of matrices and vectors. Chapter 3 through Chapter 5 give in sufficient detail a basic mathematical background necessary for state space analysis of control systems. Because modern control theory is based on theory of ordinary differential equations, a detailed treatment of the vector matrix differential equation is given in Chapter 6. Chapter 7 discusses controllability and observability of control systems. Chapter 8 presents stability analysis of linear systems and nonlinear systems. Finally, Chapter 9 presents analytical techniques for designing optimal control systems. At the end of each chapter numerous worked-out problems are provided. They supplement the development of theory and illustrate techniques presented.

The author wishes to thank Professor W.W. Seifert of MIT for his reviewing the manuscript and for his most helpful comments. The author gratefully acknowledges that a research support provided by the National Science Foundation was a major factor in the completion of this book. Finally, acknowledgment is due to the author's wife who did the typing.

KATSUHIKO OGATA
at Minneapolis, Minnesota

Contents

CHAPTER 1 INTRODUCTION 1

1-1 Introduction to State Space Concepts 1
1-2 Mathematical Background 8
1-3 State Variables 16
1-4 Optimal Control Systems 24
1-5 Classical Control Theory versus Modern Control Theory 27
Exercise Problems and Solutions 28
Problems 42

CHAPTER 2 REVIEW OF MATRICES AND VECTORS 45

2-1 Introduction 45
2-2 Rules of Matrix Operations 58
2-3 Vectors and Vector Spaces 65
2-4 Definitions of Terminologies in Matrix Analysis 76
Exercise Problems and Solutions 88
Problems 102

CHAPTER 3 COORDINATE TRANSFORMATION 106

3-1 Introduction 106
3-2 Preliminary Study—Systems with Distinct Eigenvalues 108
3-3 Diagonalization of Matrices 120
3-4 Systems with Multiple Eigenvectors—Reduction of Coefficient Matrices into the Jordan Canonical Form 132

3-5 Machine Computation of Transformation Matrices 141
3-6 Second-Order Systems 145
3-7 Systems Described by Scalar Differential Equations 146
3-8 Motions of Free Dynamic Systems 151
Exercise Problems and Solutions 155
Problems 171

CHAPTER 4 STATE SPACE REPRESENTATION OF CONTROL SYSTEMS 174

4-1 Introduction 174
4-2 State Space Representation of Systems Described by Scalar Differential Equations 175
4-3 Transforming State Space Equations into Scalar Differential Equations 196
4-4 Partial-Fraction Expansion Technique 203
4-5 State Space Equations of Systems Described by Difference Equations 218
Exercise Problems and Solutions 223
Problems 241

CHAPTER 5 MATRIX ANALYSIS 243

5-1 Introduction 243
5-2 Functions of Square Matrix 243
5-3 Sylvester's Interpolation Formula 257
5-4 Quadratic Forms and Hermitian Forms 262
5-5 Sylvester's Criteria for Definiteness and Semi-definiteness of Quadratic Forms and Hermitian Forms 277
Exercise Problems and Solutions 281
Problems 290

CHAPTER 6 OBTAINING THE SOLUTION OF THE STATE SPACE EQUATION 294

6-1 Introduction 294
6-2 Linear Vector Matrix Differential Equations 295
6-3 Solutions of Linear Time-Invariant Vector Matrix Differential Equations 307
6-4 Methods for Computing e^{At} 315
6-5 Transfer Matrices 322
6-6 Impulse Response Functions of Dynamic Systems 329
6-7 Vector Matrix Difference Equations 336
Exercise Problems and Solutions 343
Problems 367

CHAPTER 7 CONTROLLABILITY AND OBSERVABILITY 370

7-1 Introduction 370
7-2 Mathematical Background 373
7-3 Controllability of Linear Time-Invariant Systems 381
7-4 Observability of Linear Time-Invariant Systems 396
7-5 Controllability and Observability of Linear Time-Varying Systems 408
7-6 Relation between Controllability and Observability 413
Exercise Problems and Solutions 418
Problems 433

CHAPTER 8 STABILITY ANALYSIS 437

8-1 Introduction 437
8-2 Routh's and Hurwitz Stability Criteria 442
8-3 First Method of Liapunov 450
8-4 Second Method of Liapunov 452
8-5 Stability Analysis of Linear Systems 460
8-6 Stability Analysis of Nonlinear Systems 470
8-7 Stability Analysis of Discrete-Time Systems 486
8-8 Applications of Second Method of Liapunov to System Design 489
8-9 Concluding Comments 495
Exercise Problems and Solutions 496
Problems 522

CHAPTER 9 OPTIMAL CONTROL SYSTEMS 527

9-1 Introduction 527
9-2 Analytical Design of Noise-Free Optimal Regulators 529
9-3 Functional Equations Technique of Dynamic Programming 542
9-4 Analytical Design of Minimum Error Regulators 547
9-5 Minimum Error Controllers with Polynomial Inputs 558
9-6 Concluding Comments 564
Exercise Problems and Solutions 565
Problems 585

INDEX 589

1

Introduction

1-1 INTRODUCTION TO STATE SPACE CONCEPTS

The great progress in control science since about 1955 has changed basic concepts of analysis and synthesis of control systems. This progress has depended largely on mathematical study of optimal control systems.

Modern control theory which is based on state space concepts is extremely useful not only for designing a specific optimal control system but also for improving the principle on which the system will operate. By using the state space approach the control engineer may be able to design systems with performance characteristics that cannot be achieved by the classical approach by means of the frequency response method or the root locus method.

Plant and control system A physical object to be controlled is called a *plant*. It may be a heating furnace, a chemical reactor, a spacecraft, etc. A combination of components that acts together and performs a certain objective is called a *system*. A control system is a system which consists of such components as a sensor, controller, actuator, plant, etc. A plant is usually a given fixed component of a control system.

Control of complex systems The specification of the control signal over the operating time interval is called the *control law*. Control system design based on modern control theory rests on determination of an

optimal control law which minimizes a given performance index subject to certain constraints. A *performance index* is a quantitative index of performance. It is a measure of deviation from the ideal performance. Except for simple cases, optimal control of a plant requires a digital computer which utilizes feedback signals from the plant to produce the optimal control signal. Figure 1–1 shows a block diagram of a computer control system.

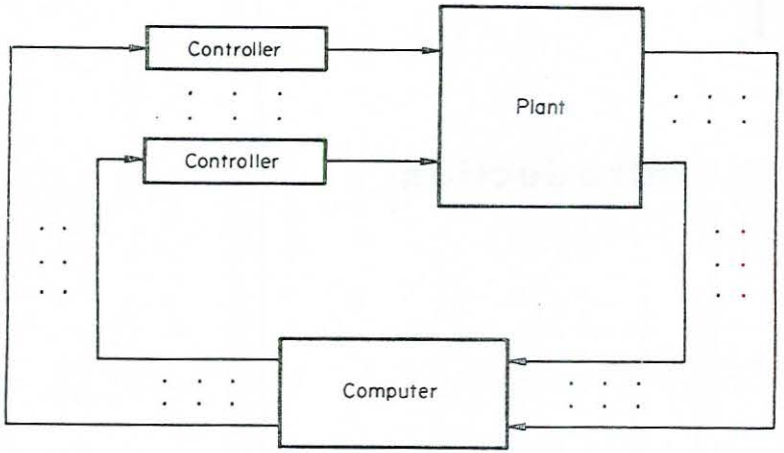


Figure 1-1 Computer control system.

The computer performs two functions: (1) producing an operational guide by use of the data from the plant; (2) computing, based on the operational guide, the command signals to the controllers which control various inputs to the plant. The command signals from the computer adjust the set points of the controllers. Computations of the command signals which are optimal are based on the mathematical model of the plant and the given performance index. In some cases, output signals of the computer are used to adjust the parameters of the controllers. (This is an adaptive control system.) The set points of the controllers are adjusted by the command signals which may be produced by the same or a different computer. Figure 1–2 shows a block diagram of such a computer control system. In either system a computer is utilized in order to minimize the given performance index, or to maximize "profit" or "efficiency," etc. of the system.

The block diagrams shown in Fig. 1–1 and Fig. 1–2 represent examples of control systems that will yield superior performance. Such control systems can be synthesized only by use of modern control theory based on state space approaches. The use of state space concepts facilitates analysis and synthesis of complex control systems in a unified manner. In synthesizing an optimal control system by using the state space technique, we need

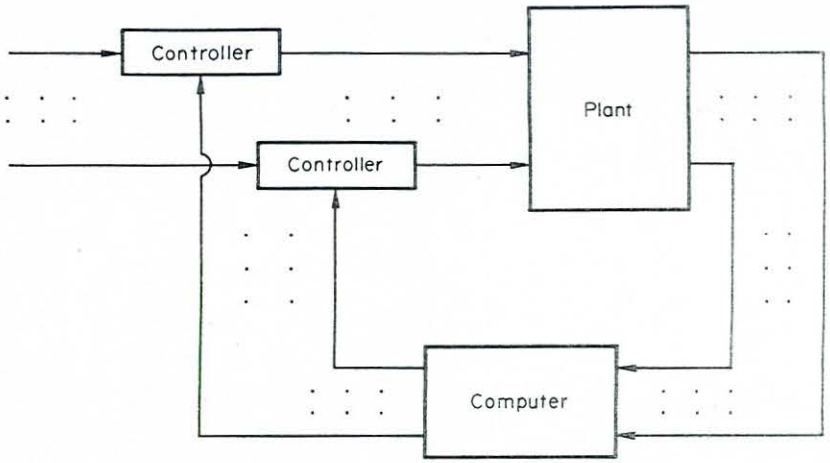


Figure 1-2 Computer control system.

solve only analytical aspects of the problem. A digital computer may be programmed to handle necessary numerical computations.

Mathematical models The first step in the analysis and synthesis of a control system is the development of a suitable mathematical model of a given plant. A mathematical model of a plant can, in principle, be obtained from the structure of the plant and the properties of the individual elements of the plant, together with fundamental physical laws. Developing a suitable mathematical model of a plant is, except in simple cases, generally a difficult problem and may lead to serious error. Difficulty in developing mathematical models arises because modern plants are, in general, extremely complex. In addition, sufficient amounts of operating data may not be available. Note that an oversimplified mathematical model may eliminate intrinsic characteristics of the plant; an overcomplicated mathematical model will cause mathematical difficulty. A suitable model is a compromise between the mathematical difficulty attached to complicated equations and the accuracy desired in the final result.

Equations describing the plant dynamics can often be approximated by ordinary differential equations of order n with time t as the independent variable. In this book we treat only systems that can be represented by nth-order ordinary differential (or difference) equations or sets of n first-order ordinary differential (or difference) equations.

An nth-order ordinary differential equation is called *linear* if each term of the equation contains at most only first power of the dependent variable or its derivatives. Systems described by linear differential equations are called *linear systems*.

Linear systems can be divided into linear time-invariant systems and linear time-varying systems. A system is time-invariant if its characteristics do not change with time. Most physical systems are time-varying. If the changes in the system's characteristics are very slow in comparison with variations in the input, a linear time-varying system can be approximated with sufficient accuracy by a linear time-invariant system.

A differential equation is called *nonlinear*, if it is not linear. Systems described by nonlinear differential equations are called *nonlinear systems*.

Any physical system, if analyzed in great detail, is nonlinear. Some physical systems can, however, be approximated with sufficient accuracy by linear systems. If the deviation from linearity is so small that it is not important for the specific problem considered, then we can treat the system as linear. Thus, whether a system can be considered linear or nonlinear depends heavily on the accuracy desired in the solution.

The difficulty of analyzing and synthesizing a system depends closely on the type of differential equation which describes it. Generally speaking, analysis and synthesis of a system become much easier if the system's differential equations are linear and time-invariant. Because such linear time-invariant differential equations represent, in an idealized sense, the behavior of many physical systems we shall treat such equations in great detail in this book.

Reduction of system equations *(1-1)*[1] A complex plant is usually characterized by multiple inputs, a few to more than ten of major importance, and a similar number of plant outputs. A reasonably accurate and complete description of the plant will require several to over a hundred first-order differential equations. Reducing these to the essential equations by hand is laborious even in the simplest case and is impractical in most cases. Without reduction, we may not be able to study a complex system on an analog computer because of difficulties with algebraic loops, even if the machine is large enough. A digital computer can be programmed to handle a complex system but even here a reduction to only the essential equations is desirable to save storage space and thus computing time. The state space formulations provide an ideal means of making this reduction in a systematic manner with no loss of essential information.

State and state space A *state* of a dynamic system is the smallest collection of numbers which must be specified at time $t = t_0$ in order to be able to predict uniquely the behavior of the system for any time $t \geqslant t_0$ for any input belonging to the given input set, provided that

[1]Numbers in parentheses refer to the list of references at the end of this chapter.

every element of the input set is known for $t \geqslant t_0$. Such numbers are called *state variables*. The *input set* here is defined as the set of all possible inputs that can be applied to the system. The state of a system at time t is uniquely determined by the state at time t_0 and the known input for $t \geqslant t_0$, and is independent of values of the state and input before t_0.

Suppose that at least n state variables $x_1, x_2, \ldots, x_n$ are needed to describe completely the behavior of a given system. Then the set of n state variables can be considered as n components of a vector $\mathbf{x}$. Such a vector $\mathbf{x}$ is called a *state vector*. A *state space* is defined as an n-dimensional space in which $x_1, x_2, \ldots, x_n$ are coordinates. The state at time t of a system defined by n first-order differential equations can then be represented by a point in an n-dimensional state space. Note that in the state space approach of analysis and synthesis of control systems we deal with a set of n first-order differential equations rather than a single nth-order differential equation.

A point in a state space at time t is called a *representative point*. The locus of representative points for a time interval $t_0 \leqslant t \leqslant t_1$ is called a trajectory for this time interval.

Review of classical theory of dynamics State space concepts, fundamental in modern control theory, evolved from the classical theory of dynamics of particles and rigid bodies. We now give a brief review of this classical theory of dynamics. *(1-2)*

In investigating dynamics of particles or rigid bodies where constraints exist, it is advantageous to specify the positions by means of independent variables rather than to denote the positions of particles or rigid bodies by rectangular coordinates, because not all rectangular coordinates are linearly independent.

Consider a system consisting of r masses, $m_1, m_2, \ldots, m_r$. Let us define coordinates of r masses as $(x_1, x_2, x_3), (x_4, x_5, x_6), \ldots, (x_{3r-2}, x_{3r-1}, x_{3r})$ and define forces acting on r masses as $(X_1, X_2, X_3), (X_4, X_5, X_6), \ldots, (X_{3r-2}, X_{3r-1}, X_{3r})$. Then Newton's equations of motion become

$$m_i \ddot{x}_i = X_i \qquad (i = 1, 2, \ldots, 3r) \tag{1-1}$$

where $m_1 = m_2 = m_3$, $m_4 = m_5 = m_6$, $\ldots$, $m_{3r-2} = m_{3r-1} = m_{3r}$. Equations (1-1) represent motions of r masses in the absence of any constraints. If there exist the following s constraint equations:

$$f_j(x_1, x_2, \ldots, x_{3r}) = 0 \qquad (j = 1, 2, \ldots, s)$$

then clearly we can define a set of $3r - s$ linearly independent variables $q_1, q_2, \ldots, q_k$ (where $k = 3r - s$) to express $x_1, x_2, \ldots, x_{3r}$, or

$$x_i = x_i(q_1, q_2, \ldots, q_k) \qquad (i = 1, 2, \ldots, 3r)$$

Such linearly independent variables $q_1, q_2, \ldots, q_k$ are called *generalized*

coordinates. The number of generalized coordinates is equal to the number of degrees of freedom of the system. In the present system the number of degrees of freedom is $k = 3r - s$. A knowledge of the magnitudes of the generalized coordinates enables us to determine the positions of the particles or rigid bodies at any given instant.

Lagrange's equations of motion of a system are differential equations in terms of $q_1, q_2, \ldots, q_k$, which are obtained by use of Hamilton's principle.[1] To simplify our discussion let us consider a conservative system without any forcing function. Lagrange showed that the equations of motion of a conservative system can be given by

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = 0 \qquad (i = 1, 2, \ldots, k) \tag{1-2}$$

where $L = T - V$

T = kinetic energy

V = potential energy

q_i = generalized coordinates

Since $L = L(q_1, q_2, \ldots, q_k; \dot{q}_1, \dot{q}_2, \ldots, \dot{q}_k)$, Eq. (1-2) reduces to a set of k second-order ordinary differential equations in $q_1, q_2, \ldots, q_k$.

Hamilton showed that the equations of motion can be simplified if we define new coordinates p_i, called the generalized momentum coordinates, such that

$$p_i = \frac{\partial T}{\partial \dot{q}_i} \qquad (i = 1, 2, \ldots, k)$$

In terms of p_i and q_i, the equations of motion can be reduced to a set of $2k$ first-order differential equations of the form:

$$\dot{q}_i = \frac{\partial H}{\partial p_i}$$
$$\dot{p}_i = -\frac{\partial H}{\partial q_i} \qquad (i = 1, 2, \ldots, k) \tag{1-3}$$

where $H = T + V$

[For derivation of Eq. (1-3), refer to Prob. A-1-4.] Equations (1-3) are called *Hamilton's equations of motion*.[2] The function H is called the *Hamiltonian function*. It is a known function of q_i and $p_i (i = 1, 2, \ldots, k)$.

For a system with k degrees of freedom a $2k$-dimensional space in which $q_1, q_2, \ldots, q_k$ and $p_1, p_2, \ldots, p_k$ are coordinates is called a *phase space*. A representative point in the phase space moves about the phase space, as both coordinates and momenta change with time. Hamilton's equations (1-3) give the components of the velocity of the representative point in the phase space. The problem of classical dynamics of particles and rigid bodies is to investigate the path of the representative point in the phase space.

Notice that Eq. (1-3) is a set of $2k$ first-order differential equations of the following form:

$$\dot{q}_i = f_i(q_1, q_2, \ldots, q_k; p_1, p_2, \ldots, p_k)$$
$$\dot{p}_i = g_i(q_1, q_2, \ldots, q_k; p_1, p_2, \ldots, p_k) \tag{1-4}$$
$$(i = 1, 2, \ldots, k)$$

In the system of Eq. (1-4) the $2k$ variables $q_1, q_2, \ldots, q_k$; $p_1, p_2, \ldots, p_k$ describe completely the dynamics of the system. Since these $2k$ variables are the smallest collection of numbers which must be specified in order to describe the system dynamics completely, they are state variables.

As an example consider the simple pendulum shown in Fig. 1-3. In considering this system it is not convenient to use ordinary rectangular coordinates x and y, because there is a constraint equation, $x^2 + y^2 = l^2$. It is convenient to choose polar coordinates θ and r, in which r is constant. Thus the only variable of the system or the generalized coordinate is θ.

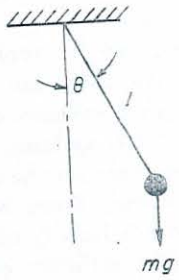


Figure 1-3 Simple pendulum.

In this system the mass m has risen from its equilibrium position a distance $l(1 - \cos\theta)$. The potential energy is therefore given by

$$V = mgl(1 - \cos\theta)$$

The velocity of the mass m is $l\dot{\theta}$. Hence, the kinetic energy is given by

$$T = \frac{m}{2}(l\dot{\theta})^2$$

Let us put $\theta = q$. The Hamiltonian function H is given by

$$H = T + V = \frac{m}{2}(l\dot{q})^2 + mgl(1 - \cos q)$$

Since p is defined by

$$p = \frac{\partial T}{\partial \dot{q}} = ml^2\dot{q}$$

H can be written as

$$H = \frac{1}{2ml^2}p^2 + mgl(1 - \cos q)$$

From Eq. (1-3)

$$\dot{q} = \frac{\partial H}{\partial p} = \frac{1}{ml^2}p$$
$$\dot{p} = -\frac{\partial H}{\partial q} = -mgl\sin q \tag{1-5}$$

[1] See Prob. A-1-1.

[2] In solving problems of vibratory systems, Lagrange's equations are more convenient than Hamilton's. For theoretical investigation of dynamic systems, however, the latter are more convenient.

Equations (1–5) define the motion of the pendulum in the phase plane in which q and p are the coordinates. Here q and p completely describe the dynamics of this system. They are a set of state variables.

Phase spaces have been used in the classical theory of dynamics to describe dynamics of systems consisting of particles or rigid bodies. For a system with k degrees of freedom the phase space is a $2k$-dimensional space in which $q_1, q_2, \ldots, q_k$ and $p_1, p_2, \ldots, p_k$ are coordinates. Hence for a one-degree-of-freedom system, the phase space becomes a phase plane, a two-dimensional phase space. For a two-degree-of-freedom system, the phase space is four-dimensional.

The concept of degrees of freedom as used in the classical theory of dynamics may not apply to control systems. Hence the definition of the phase space previously given may not be directly applicable to control systems. Concepts of phase space can, however, be directly applied to control systems. In fact, concepts of state space are evolution of concepts of phase space. Note that definitions of the state space used in modern control theory are more general than the definition of the phase space used in the classical theory of dynamics. (The "three-dimensional phase space" frequently used in the analysis and synthesis of optimal control systems actually means a three-dimensional state space.)

This completes our brief review of phase space concepts used in the classical theory of dynamics of particles and rigid bodies.

In the next section, we present a mathematical background necessary for subsequent discussions on state variables and state spaces.

1-2 MATHEMATICAL BACKGROUND

In this section we shall first discuss representation of system differential equations in terms of vector matrix notation. We shall next review existence and uniqueness theorems concerning solutions of differential equations. These theorems form a basis for precisely assigning a set of state variables to a given system with a given input set.

Representation of system differential equations in terms of vector matrix notation Systems which we shall consider in this book are assumed to be described by ordinary differential equations of order n, or by a set of n first-order ordinary differential equations.

We shall first show that a differential equation of order n can always be reduced to a set of n first-order differential equations. (We shall give detailed discussions on this subject in Chapter 4.) Consider the following differential equation of order n:

$$\overset{(n)}{x} = f(x, \dot{x}, \ldots, \overset{(n-1)}{x}, u, t) \tag{1–6}$$

where u is the forcing function. By defining new variables $x_1, x_2, \ldots, x_n$ such that

$$\begin{aligned} x_1 &= x \\ x_2 &= \dot{x} \\ &\cdots \\ x_n &= \overset{(n-1)}{x} \end{aligned}$$

Eq. (1–6) can be reduced to

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 \\ &\cdots \\ \dot{x}_{n-1} &= x_n \\ \dot{x}_n &= f(x_1, x_2, \ldots, x_n, u, t) \end{aligned} \tag{1–7}$$

Equations (1–7) are a set of n first-order differential equations.

We shall next consider a more general case. Consider the control system shown in Fig. 1–4. Equations of the plant may be adequately described by a set of n first-order ordinary differential equations of the following form:

$$\dot{x}_i = f_i(x_1, x_2, \ldots, x_n; u_1, u_2, \ldots, u_r; t) \tag{1–8}$$
$$(i = 1, 2, \ldots, n)$$

where $x_1, x_2, \ldots, x_n$ are state variables, and $u_1, u_2, \ldots, u_r$ are control variables. It is noted that the state variables x_i are not necessarily measurable. We may have to measure something else (usually, system outputs) and

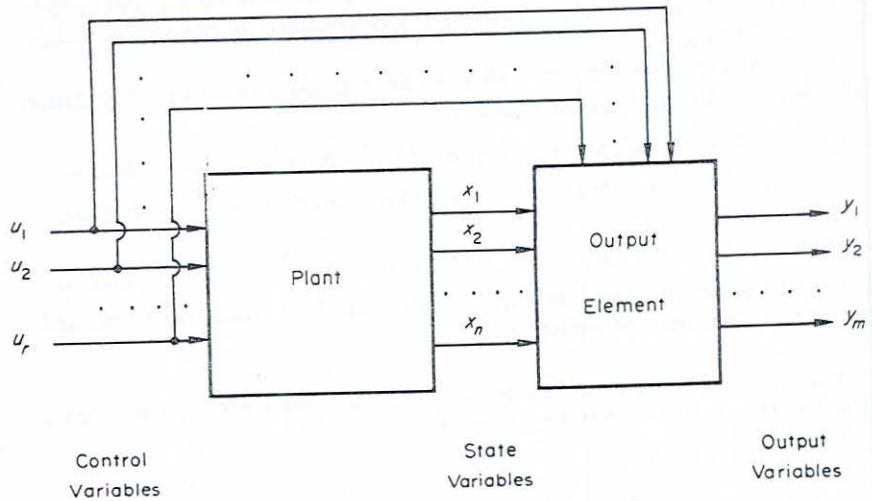


Figure 1-4 Block diagram of control system, showing terminology.

then estimate what the actual state variables are. We assume that the system outputs $y_1, y_2, \ldots, y_m$ are directly measurable and are related to the state variables $x_1, x_2, \ldots, x_n$ and control variables $u_1, u_2, \ldots, u_r$ by the following equations:

$$y_j = g_j(x_1, x_2, \ldots, x_n; u_1, u_2, \ldots, u_r; t) \qquad (1\text{–}9)$$
$$(j = 1, 2, \ldots, m; m \leqslant n)$$

If the order of the system is other than small, such a set of equations becomes notationally complicated. It becomes necessary to use vectors and matrices to simplify the notation.[1] Rewritten with vector notation, Eq. (1–8) and Eq. (1–9) become respectively

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \qquad (1\text{–}10)$$

and

$$\mathbf{y} = \mathbf{g}(\mathbf{x}, \mathbf{u}, t)$$

where $\mathbf{x}$, $\mathbf{u}$, $\mathbf{y}$, $\mathbf{f}(\mathbf{x}, \mathbf{u}, t)$,and $\mathbf{g}(\mathbf{x}, \mathbf{u}, t)$ are vectors defined by

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}, \qquad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ \cdot \\ u_r \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_m \end{bmatrix}$$

$$\mathbf{f}(\mathbf{x}, \mathbf{u}, t) = \begin{bmatrix} f_1(\mathbf{x}, \mathbf{u}, t) \\ f_2(\mathbf{x}, \mathbf{u}, t) \\ \cdot \\ \cdot \\ \cdot \\ f_n(\mathbf{x}, \mathbf{u}, t) \end{bmatrix}, \qquad \mathbf{g}(\mathbf{x}, \mathbf{u}, t) = \begin{bmatrix} g_1(\mathbf{x}, \mathbf{u}, t) \\ g_2(\mathbf{x}, \mathbf{u}, t) \\ \cdot \\ \cdot \\ \cdot \\ g_m(\mathbf{x}, \mathbf{u}, t) \end{bmatrix}$$

Here vectors $\mathbf{x}$, $\mathbf{u}$, and $\mathbf{y}$ are called respectively the *state* vector, the *control* vector, and the *output* vector. The system is specified by the vector-valued functions $\mathbf{f}$ and $\mathbf{g}$.

If the system shown in Fig. 1–4 is linear in $\mathbf{x}$ and $\mathbf{u}$, a set of n first-order differential equations may be written as

$$\begin{aligned}
\dot{x}_1 &= a_{11}(t)x_1 + a_{12}(t)x_2 + \cdots + a_{1n}(t)x_n + b_{11}(t)u_1 + \cdots + b_{1r}(t)u_r \\
\dot{x}_2 &= a_{21}(t)x_1 + a_{22}(t)x_2 + \cdots + a_{2n}(t)x_n + b_{21}(t)u_1 + \cdots + b_{2r}(t)u_r \\
&\cdots \\
\dot{x}_n &= a_{n1}(t)x_1 + a_{n2}(t)x_2 + \cdots + a_{nn}(t)x_n + b_{n1}(t)u_1 + \cdots + b_{nr}(t)u_r
\end{aligned}$$

and m algebraic equations relating output variables, state variables, and control variables may be written as

[1] In this book, lower case boldface Roman or Greek letters denote vectors. Capital boldface Roman or Greek letters denote matrices.

$$\begin{aligned}
y_1 &= c_{11}(t)x_1 + c_{12}(t)x_2 + \cdots + c_{1n}(t)x_n + d_{11}(t)u_1 + \cdots + d_{1r}(t)u_r \\
y_2 &= c_{21}(t)x_1 + c_{22}(t)x_2 + \cdots + c_{2n}(t)x_n + d_{21}(t)u_1 + \cdots + d_{2r}(t)u_r \\
&\cdots \\
y_m &= c_{m1}(t)x_1 + c_{m2}(t)x_2 + \cdots + c_{mn}(t)x_n + d_{m1}(t)u_1 + \cdots + d_{mr}(t)u_r
\end{aligned}$$

Rewritten in terms of vector matrix notation

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u}$$
$$\mathbf{y} = \mathbf{C}(t)\mathbf{x} + \mathbf{D}(t)\mathbf{u}$$

where

$$\mathbf{A}(t) = \begin{bmatrix} a_{11}(t) & a_{12}(t) & \cdots & a_{1n}(t) \\ a_{21}(t) & a_{22}(t) & \cdots & a_{2n}(t) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n1}(t) & a_{n2}(t) & \cdots & a_{nn}(t) \end{bmatrix}$$

$$\mathbf{B}(t) = \begin{bmatrix} b_{11}(t) & b_{12}(t) & \cdots & b_{1r}(t) \\ b_{21}(t) & b_{22}(t) & \cdots & b_{2r}(t) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ b_{n1}(t) & b_{n2}(t) & \cdots & b_{nr}(t) \end{bmatrix}$$

$$\mathbf{C}(t) = \begin{bmatrix} c_{11}(t) & c_{12}(t) & \cdots & c_{1n}(t) \\ c_{21}(t) & c_{22}(t) & \cdots & c_{2n}(t) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ c_{m1}(t) & c_{m2}(t) & \cdots & c_{mn}(t) \end{bmatrix}$$

$$\mathbf{D}(t) = \begin{bmatrix} d_{11}(t) & d_{12}(t) & \cdots & d_{1r}(t) \\ d_{21}(t) & d_{22}(t) & \cdots & d_{2r}(t) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ d_{m1}(t) & d_{m2}(t) & \cdots & d_{mr}(t) \end{bmatrix}$$

The elements of $\mathbf{A}(t)$ depend on the plant parameters, and the elements of $\mathbf{B}(t)$ show how each control variable affects the state variables of the plant. $\mathbf{C}(t)$ shows how the output variables are related to the state variables and $\mathbf{D}(t)$ shows how the control variables directly affect the output variables. Here in any practical case $m \leqslant n$.

We have presented a way to describe system equations in terms of vectors and matrices. For theoretical work, notational simplicity gained by vector matrix operation is most convenient and is actually essential for state space analysis of control systems. Such vector matrix representation of the

system equation in which x is a state vector is called a *state space representation* of the system equation. Large, complex problems can be handled with ease by the systematic format of the state space representation.

Example 1-1 Let us obtain a state space representation of the following linear time-invariant system:

$$\frac{d^n x}{dt^n} + a_1 \frac{d^{n-1} x}{dt^{n-1}} + \cdots + a_{n-1} \frac{dx}{dt} + a_n x = 0 \qquad (1\text{–}11)$$

Let us put

$$\begin{aligned} x_1 &= x \\ x_2 &= \dot{x} \\ &\cdots \\ x_n &= x^{(n-1)} \end{aligned}$$

Equation (1–11) is equivalent to the following n simultaneous equations:

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 \\ &\cdots \\ \dot{x}_{n-1} &= x_n \\ \dot{x}_n &= -a_n x_1 - \cdots - a_1 x_n \end{aligned}$$

Hence Eq. (1–11) can be written as the following state space equation:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_1 \end{bmatrix}$$

Existence and uniqueness of solutions of differential equations We shall next review existence and uniqueness theorems concerning solutions of ordinary differential equations. Consider a free dynamic system defined by the following first-order differential equation:

$$\dot{x} = f(x, t)$$

$$x(t_0) = x_0$$

Let us find the solution of this equation. Geometrically, the solution is an integral curve passing through point $P_0(x_0, t_0)$. Since the slope of the integral curve at point P_0 is $f(x_0, t_0)$, it is possible to draw a short line segment from point P_0 with the slope $f(x_0, t_0)$. Let us call the end point of the short line segment $P_1(x_1, t_1)$. Then P_0P_1 is an approximation to the integral curve

in a neighborhood of point P_0. Next, draw a short line segment from point P_1 with the slope $f(x_1, t_1)$. Call the end point of the short line segment $P_2(x_2, t_2)$. Repeating this process, we can construct a piecewise linear curve $P_0P_1P_2\ldots$ consisting of short line segments, as shown in Fig. 1–5. If we make the length of each line segment smaller and smaller by increasing the number of line segments, then in the limit $P_0P_1P_2\ldots$ will converge to a curve.

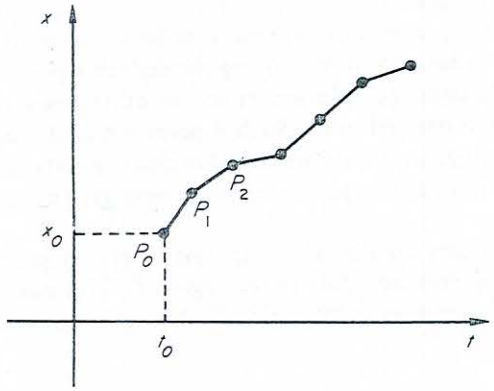


Figure 1-5 Integral curve, showing solution of $\dot{x} = f(x, t)$.

We may intuitively expect that the limiting curve will be the integral curve of the given differential equation. In order for this to be true, however, there are certain conditions which $f(x, t)$ must satisfy. It is a well-known result in theory of ordinary differential equations that such conditions are given as the Cauchy-Lipschitz theorem.

Theorem 1-1 Cauchy-Lipschitz theorem In a dynamic system defined by

$$\dot{x} = f(x, t), \qquad x(t_0) = x_0 \qquad (1\text{–}12)$$

suppose that

(a) $f(x, t)$ is a single-valued, continuous function in the region Γ defined by

$$|x - x_0| < h, \qquad 0 \leqslant t - t_0 < \tau$$

and there exists a positive number M such that

$$|f(x, t)| \leqslant M$$

in the region Γ, and

(b) the following Lipschitz condition

$$|f(x, t) - f(z, t)| \leqslant K|x - z|$$

where points (x, t) and (z, t) are arbitrary points in the region Γ and K is a positive constant, is satisfied.

Then there exists a unique solution defined in the interval

$$0 \leqslant t - t_0 < T, \qquad T = \min\left(\tau, \frac{h}{M}\right)$$

and satisfying the initial condition $x(t_0) = x_0$. The solution is continuous in the interval $0 \leqslant t - t_0 < T$. The curve $P_0P_1P_2\ldots$ converges to the integral curve.

We shall not give a proof of this theorem here, but it is given in Prob. A-1-11. In this theorem, the notation min $[\tau, (h/M)]$ means that we take the minimum of the values of τ and h/M. If h is ∞, then the solution $x = x(t)$ is unique in the interval $0 \leqslant t - t_0 < T = \tau$.

Note that if $f(x, t)$ does not possess a definite value at (x_0, t_0), then the existence of a unique solution passing through this point (x_0, t_0) is not guaranteed. In some cases no solution exists; in other cases there exists a solution or more than one solution. Such a point is called a *singular* point. Points that are not singular points are called *ordinary* points. In the absence of any forcing function, the trajectory passing through any ordinary point is unique.

Notice that Lipschitz condition is satisfied if $f(x, t)$ possesses a continuous first partial derivative $\partial f/\partial x$ in the region Γ. This can be seen easily because if $\partial f/\partial x$ is continuous in the region Γ, there exists a positive constant K such that

$$\left|\frac{\partial f}{\partial x}\right| \leqslant K$$

Hence

$$|f(x, t) - f(z, t)| = \left|\int_z^x \frac{\partial f}{\partial x}dx\right| \leqslant K|x - z|$$

Let us next consider a more general case where an nth-order free dynamic system is described by

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t), \qquad \mathbf{x}(t_0) = \mathbf{x}_0$$

or

$$\begin{aligned}
\dot{x}_1 &= f_1(x_1(t), \ldots, x_n(t), t) \\
\dot{x}_2 &= f_2(x_1(t), \ldots, x_n(t), t) \\
&\cdots \quad \cdots \quad \cdots \\
\dot{x}_n &= f_n(x_1(t), \ldots, x_n(t), t)
\end{aligned} \tag{1–13}$$

where $x_i(t_0) = x_{i0} \qquad (i = 1, 2, \ldots, n)$

For this case, the Cauchy-Lipschitz theorem can be stated as follows:

Theorem 1-2 Suppose that

(a) the functions $f_i(x_1, x_2, \ldots, x_n, t)$ in Eq. (1–13) are single-valued and continuous in the region Γ defined by

$$|x_i - x_{i0}| < h_i \qquad (i = 1, 2, \ldots, n)$$
$$0 \leqslant t - t_0 < \tau$$

and there exist positive numbers M_i $(i = 1, 2, \ldots, n)$ such that

$$|f_i(x_1, x_2, \ldots, x_n, t)| \leqslant M_i$$

in the region Γ and

(b) the following Lipschitz conditions:

$$|f_i(x_1, x_2, \ldots, x_n, t) - f_i(z_1, z_2, \ldots, z_n, t)| \leqslant K\sum_{j=1}^{n}|x_j - z_j|,$$
$$(i = 1, 2, \ldots, n)$$

where points $(x_1, x_2, \ldots, x_n, t)$ and $(z_1, z_2, \ldots, z_n, t)$ are arbitrary points in the region Γ and K is a positive constant, are satisfied.

Then there exists a unique solution defined in the interval

$$0 \leqslant t - t_0 < T, \qquad T = \min\left(\tau, \frac{h_1}{M_1}, \cdots, \frac{h_n}{M_n}\right)$$

and satisfying the initial conditions $x_i(t_0) = x_{i0}$ $(i = 1, 2, \ldots, n)$.

We shall not give a proof of this theorem here. Instead we refer the reader to Prob. A-1-11 and Prob. A-1-12.

Suppose that the assumptions made in Theorem 1–2 are satisfied. We obtain a set of unique solutions $x_1(t), x_2(t), \ldots, x_n(t)$ of Eq. (1–13) for every set of the initial conditions $x_{10}, x_{20}, \ldots, x_{n0}$. Hence the solution of the system of Eq. (1–13) involves $x_{10}, x_{20}, \ldots, x_{n0}$. By putting

$$x_{i0} = c_i \qquad (i = 1, 2, \ldots, n)$$

the solution can be written as

$$\begin{aligned}
x_1 &= x_1(c_1, c_2, \ldots, c_n, t) \\
x_2 &= x_2(c_1, c_2, \ldots, c_n, t) \\
&\cdots \quad \cdots \quad \cdots \\
x_n &= x_n(c_1, c_2, \ldots, c_n, t)
\end{aligned}$$

The solution involving n arbitrary constants is called a *general* solution of the nth-order system.

In the next section we shall discuss ways of choosing a set of state variables for a given system with a given input set. There we shall utilize the result obtained in this section that in Eq. (1–13) if the conditions specified in Theorem 1–2 are satisfied, then $x_1(t), x_2(t), \ldots, x_n(t)$ for $0 \leqslant t - t_0 < T$ can be determined uniquely for the given initial conditions, $x_1(t_0), x_2(t_0), \ldots, x_n(t_0)$.

1-3 STATE VARIABLES

In this section we shall discuss what variables in the system can constitute a set of state variables for the system once the input set, an aggregate of all the input functions under consideration, is given. The general procedure for associating a set of state variables or a state vector with the system is to associate with the system a vector $\mathbf{x}(t_0)$ such that $\mathbf{x}(t)$ is uniquely determined by $\mathbf{x}(t_0)$ and $\mathbf{u}(t)$. In the following discussions we assume that the elements of $\mathbf{x}(t)$ are linearly independent of each other.

Based on Theorem 1–2 presented in Section 1–2 we can make the following statement: In the linear or nonlinear system defined by Eq. (1–10) rewritten as

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, t)$$

if for any input vector $\mathbf{u}(t)$ which belongs to the input set considered for the system, the vector function $\mathbf{f}(\mathbf{x}, \mathbf{u}, t)$ is a single-valued vector function of its arguments and is sufficiently smooth in the time interval $0 \leqslant t - t_0 < T$, then the system possesses a unique solution starting at any $\mathbf{x}(t_0)$ and subjected to the given $\mathbf{u}(t)$. In such a system $\mathbf{x}(t)$ is clearly a state vector associated with the system, since $\mathbf{x}(t)$ can be uniquely determined for a given $\mathbf{x}(t_0)$ and any element $\mathbf{u}(t)$ of the input set. Then n components of the vector $\mathbf{x}$, namely $x_1, x_2, \ldots, x_n$ are a set of state variables. The unique solution of this system observed at time t after starting at time t_0 in state $\mathbf{x}(t_0) = \mathbf{x}_0$ and affected by the input vector $\mathbf{u}(t)$ is sometimes written as $\mathbf{x}(t) = \boldsymbol{\phi}_u(t; \mathbf{x}_0, t_0)$. This notation clearly indicates that the present state vector $\mathbf{x}(t)$ is a function of the input vector $\mathbf{u}(t)$ and the initial state $\mathbf{x}_0$.

Note that in the forced dynamic system, if the input vector $\mathbf{u}(t)$ exhibits a jump at $t = t_0$, then the initial state (at $t = t_0$) becomes ambiguous. To avoid this inconvenience we shall assume, unless otherwise stated, that the initial state is established at $t = t_0-$ as a point in the state space and the input (which may include discontinuities) is applied at $t = t_0$. Therefore, if the input is an impulse occurring at $t = t_0$, then the initial state is established at $t = t_0-$ and a jump in the state occurs at $t = t_0$.

Example 1-2 Consider the motion of a rigid body which has one point fixed at O. The motion is given by the following set of three first-order differential equations:

$$\begin{aligned} A\dot{\omega}_1 &= (B - C)\omega_2\omega_3 + u_1 \\ B\dot{\omega}_2 &= (C - A)\omega_3\omega_1 + u_2 \\ C\dot{\omega}_3 &= (A - B)\omega_1\omega_2 + u_3 \end{aligned}$$

where A, B, and C are the moments of inertia about the principal axes Ox, Oy, and Oz, respectively; ω_1, ω_2, and ω_3 are their component angular velocities; u_1, u_2, and u_3 are the known torque inputs about the axes Ox, Oy, and Oz, respectively. In this system ω_1, ω_2, and ω_3 are a set of state variables.

Linear systems Since linear systems are our main concern in this book, we shall pay a particular attention to such systems. Consider the linear system defined by

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u} \tag{1–14}$$

or

$$\begin{aligned} \dot{x}_1 &= a_{11}(t)x_1 + a_{12}(t)x_2 + \cdots + a_{1n}(t)x_n + p_1(t) \\ \dot{x}_2 &= a_{21}(t)x_1 + a_{22}(t)x_2 + \cdots + a_{2n}(t)x_n + p_2(t) \\ &\cdots \quad \cdots \quad \cdots \\ \dot{x}_n &= a_{n1}(t)x_1 + a_{n2}(t)x_2 + \cdots + a_{nn}(t)x_n + p_n(t) \end{aligned}$$

where

$$\begin{aligned} p_1(t) &= b_{11}(t)u_1 + b_{12}(t)u_2 + \cdots + b_{1r}(t)u_r \\ p_2(t) &= b_{21}(t)u_1 + b_{22}(t)u_2 + \cdots + b_{2r}(t)u_r \\ &\cdots \quad \cdots \quad \cdots \\ p_n(t) &= b_{n1}(t)u_1 + b_{n2}(t)u_2 + \cdots + b_{nr}(t)u_r \end{aligned} \tag{1–15}$$

The matrices $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are given, and $\mathbf{u}$, which belongs to a given input set, is a known vector for the time interval $0 \leqslant t - t_0 < T$. Therefore, $p_1(t), p_2(t), \ldots, p_n(t)$ are known functions of time for the time interval $0 \leqslant t - t_0 < T$. Theorem 1–2 states that if the $a_{ij}(t)$ and $p_i(t)$ $(i, j = 1, 2, \ldots, n)$ are continuous functions of time in the time interval $0 \leqslant t - t_0 < T$, then Eq. (1–14) possesses a unique solution satisfying the initial conditions. Hence if the $a_{ij}(t)$ and $b_{ik}(t)$ $(i, j = 1, 2, \ldots, n;\ k = 1, 2, \ldots, r)$ are continuous functions of time and the $u_k(t)$ are known continuous functions of time in the interval $0 \leqslant t - t_0 < T$, then $x_1(t), x_2(t), \ldots, x_n(t)$ are uniquely determined by $x_1(t_0), x_2(t_0), \ldots, x_n(t_0)$ and $u_1(t), u_2(t), \ldots, u_r(t)$. Therefore, if this uniqueness holds for any element of the input set, then $x_1, x_2, \ldots, x_n$ are a set of state variables for the system for the given input set.

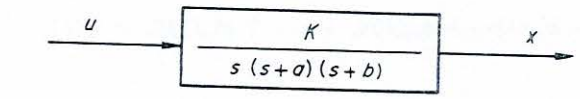
Suppose next that the $p_i(t)$ are not necessarily continuous in the time interval $0 \leqslant t - t_0 < T$ (where T can be infinity), but may have discontinuities at $t = t_k$ $(k = 0, 1, 2, \ldots)$. Then the variables $x_1, x_2, \ldots, x_n$ can still be a set of state variables for the given input set, if at each discontinuity $t = t_k (k = 0, 1, 2, \ldots)$ the initial condition for a continuous time interval $t_{k+1} \leqslant t < t_{k+2}$ can be uniquely determined from the end condition for the previous time interval $t_k \leqslant t < t_{k+1}$ and if this holds true for any element $\mathbf{u}$ of the given input set. The reason for this is that since $x_1 x_2, \ldots, x_n$ are a set of state variables for each continuous time interval $t_k \leqslant t < t_{k+1}$, if the amount of jump, if any, of the trajectory at each discontinuity can be uniquely determined, then the solution for the time interval $0 \leqslant t - t_0 < T$ satisfying the initial conditions $x_i(t_0), x_2(t_0), \ldots, x_n(t_0)$ exists and is unique for a given $\mathbf{u}$. Since Eq. (1–14) is a first-order differential equation, the

amount of jump in the trajectory is finite for the $p_i(t)$ involving up to the impulse function, where the unit impulse function or the delta function is defined by

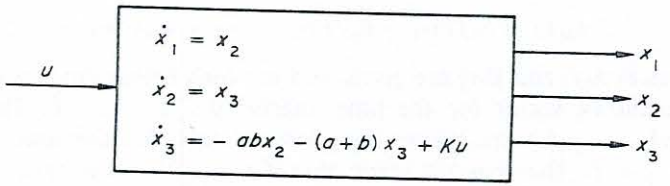
$$\delta(t) = \infty \qquad \text{for } t = 0$$
$$\delta(t) = 0 \qquad \text{for } t \neq 0$$
$$\int_{-\infty}^{\infty} \delta(t)\, dt = 1$$

Example 1-3 Consider the plant shown in Fig. 1-6(A). It may be an electric motor. The differential equation of the plant is given by

$$\ddot{x} + (a + b)\ddot{x} + ab\dot{x} = Ku \tag{1-16}$$



(A)



(B)

Figure 1-6 Block diagrams for a plant. (A) Transfer function representation; (B) state space representation, showing state variables and control function.

This third-order differential equation can be transformed into a set of three first-order differential equations by defining a set of variables. For example, let us define

$$x_1 = x$$
$$x_2 = \dot{x}$$
$$x_3 = \ddot{x}$$

then Eq. (1-16) can be written as

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = x_3$$
$$\dot{x}_3 = -abx_2 - (a + b)x_3 + Ku$$

In this case, x_1, x_2, and x_3 are a set of state variables. If the plant represents an electric motor, then x_1 denotes the angular position of the output shaft, x_2 the angular velocity of the output shaft, and x_3 the angular acceleration of the output shaft. The control function or the input to the plant, $u(t)$, corresponds to the input voltage used to drive the electric motor. Figure 1-6(B) shows a state space representation of the plant shown in Fig. 1-6(A).

Note that, in general, it is desirable that the state variables be quantities directly and accurately measurable. This, however, is not necessarily possible in a given situation because of the presence of noise.

In the use of relay control, numerator dynamics in the transfer function of a plant would give rise to discontinuities in the output variable and/or its derivatives. This problem can be treated by introducing new coordinates to handle the discontinuous variables. We shall consider next a case where discontinuities arise in the derivatives of the output.

If any of the $p_i(t)$ of Eq. (1-15) involves a doublet function[1], triplet function,[2] etc., then the amount of a jump of the trajectory in the state space at time of discontinuity may become infinity and the uniqueness of the solution may be lost. Hence, in this case $x_1, x_2, \ldots, x_n$ cannot be a set of state variables. In such a case, it is necessary to choose an appropriate set of variables as state variables. We proceed to discuss this case. (In Chapter 4 we present more thorough discussions of such a case.) Consider the plant shown in Fig. 1-7. This plant involves numerator dynamics. The differential equation for this plant is given by

$$\overset{(n)}{x} + a_1\overset{(n-1)}{x} + \cdots + a_n x = b_0\overset{(m)}{u} + b_1\overset{(m-1)}{u} + \cdots + b_m u \tag{1-17}$$
$$1 \leqslant m \leqslant n$$

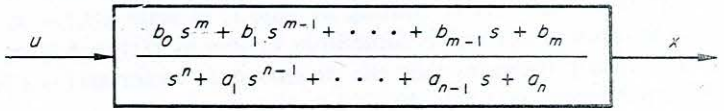


Figure 1-7 Block diagram of plant with numerator dynamics.

Suppose that the input function u is a discontinuous function of time t possibly involving the delta function, but does not involve any higher-order impulse functions, such as a doublet function, triplet function, etc. If we define

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} x \\ \dot{x} \\ \cdot \\ \cdot \\ \cdot \\ \overset{(n-1)}{x} \end{bmatrix} \tag{1-18}$$

[1]The *unit doublet function* is defined as the derivative of the unit impulse function.
[2]The *unit triplet function* is defined as the derivative of the unit doublet function.

then Eq. (1–17) becomes

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 \\ &\cdots \\ \dot{x}_{n-1} &= x_n \\ \dot{x}_n &= -a_n x_1 - \cdots - a_1 x_n + b_0 \overset{(m)}{u} + \cdots + b_m u \end{aligned} \tag{1–19}$$

Notice that in this case if u possesses a finite discontinuity at $t = t_1$, then $\dot{u}$ will possess the delta function at $t = t_1$, and higher derivatives of u possess higher-order impulse functions, such as a doublet function, triplet function, etc. at $t = t_1$. If the right-hand side of the last equation in Eq. (1–19) involves the doublet function or higher impulse functions, then the amount of the jump in the trajectory at the time of discontinuity can become infinity. This means that the state $\mathbf{x}(t)$ after the discontinuity at $t = t_1$ cannot be determined uniquely by the state just before the discontinuity at $t = t_1$, even though the input $u(t)$ is known. Therefore, $x_1, x_2, \ldots, x_n$ defined by Eq. (1–18) cannot be a set of state variables.

In such a case we can choose a set of n variables consisting of suitable combinations of derivatives of various orders of $x(t)$ and $u(t)$ such that the resulting trajectories will not exhibit infinite jumps in the new state space. (Note that finite jumps in the trajectory do not bother us, provided that we can compute the amounts of jumps. The amounts of jumps in the trajectory depend on the state space chosen.)

Example 1-4 Suppose that the plant shown in Fig. 1–7 is given by

$$\frac{X(s)}{U(s)} = \frac{s^3 + 8s^2 + 17s + 8}{s^3 + 6s^2 + 11s + 6}$$

or

$$\dddot{x} + 6\ddot{x} + 11\dot{x} + 6x = \dddot{u} + 8\ddot{u} + 17\dot{u} + 8u \tag{1–20}$$

Assume that the input set in which u is an element consists of all functions of time t involving up to the delta function, but not higher-order impulse functions, such as a doublet function, triplet function, etc. In this case x, $\dot{x}$, and $\ddot{x}$ cannot constitute a set of state variables. If we put

$$\begin{aligned} x_1 &= x - u \\ x_2 &= \dot{x} - \dot{u} + 6x - 8u \\ x_3 &= \ddot{x} - \ddot{u} + 6\dot{x} - 8\dot{u} + 11x - 17u \end{aligned}$$

then Eq. (1–20) can be written as

$$\begin{aligned} \dot{x}_1 &= -6x_1 + x_2 + 2u \\ \dot{x}_2 &= -11x_1 + x_3 + 6u \\ \dot{x}_3 &= -6x_1 + 2u \end{aligned} \tag{1–21}$$

In the space, whose coordinates are x_1, x_2, and x_3, the trajectory of the motion starting from any initial condition point can be determined uniquely for a given

input u which belongs to the input set considered. Hence, x_1, x_2, and x_3 are a set of state variables.

Note that by solving Eq. (1–21) we can obtain x_1 as a function of t. Then $x = x_1 + u$, the solution to Eq. (1–20), can be determined as a function of t, since u is a known function of t.

We can summarize what we have previously stated in regard to the linear system as follows:

(i) Suppose that in the linear system defined by Eq. (1–14), rewritten as

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u}$$

the elements of $\mathbf{A}(t)$ are continuous functions of time in the time interval $0 \leqslant t - t_0 < T$. If the elements of $\mathbf{B}(t)\mathbf{u}$ are integrable (namely, the elements of $\int_{t_0}^{T+t_0} \mathbf{B}(t)\mathbf{u}\, dt$ assume definite values) for all $\mathbf{u}$ in the input set, then $\mathbf{x}$ is a state vector for the system for the given input set defined for the time interval $0 \leqslant t - t_0 < T$. The space having $x_1, x_2, \ldots, x_n$ as the coordinates is a state space.

If, however, any of the elements of $\mathbf{B}(t)\mathbf{u}$ is not integrable (namely if any element of $\int_{t_0}^{T+t_0} \mathbf{B}(t)\mathbf{u}\, dt$ becomes infinity or does not possess a unique value), then $\mathbf{x}$ cannot be a state vector. Hence, if any element of $\mathbf{B}(t)\mathbf{u}$ involves a doublet function, triplet function, etc., or if any element of $\mathbf{B}(t)\mathbf{u}$ is not single-valued, then $\mathbf{x}$ cannot be a state vector.

(ii) In the system defined by

$$\overset{(n)}{x} + a_1(t)\overset{(n-1)}{x} + \cdots + a_n(t)x = b_0(t)\overset{(m)}{u} + b_1(t)\overset{(m-1)}{u} + \cdots + b_m(t)u$$
$$1 \leqslant m \leqslant n$$

or

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 \\ &\cdots \\ \dot{x}_{n-1} &= x_n \\ \dot{x}_n &= -a_n(t)x_1 - \cdots - a_1(t)x_n + p_n(t) \end{aligned}$$

where

$$\begin{aligned} x_1 &= x \\ p_n(t) &= b_0(t)\overset{(m)}{u} + b_1(t)\overset{(m-1)}{u} + \cdots + b_m(t)u \end{aligned}$$

if the forcing function u is not sufficiently smooth that $p_n(t)$ involves higher-order impulse functions, such as a doublet function, triplet function, etc., then $x_1, x_2, \ldots, x_n$ cannot be a set of state variables. In such a case a suitable choice of n variables may be made so that the system equation can be modified to

$$\dot{\mathbf{y}} = \mathbf{C}(t)\mathbf{y} + \mathbf{D}(t)u$$

where no derivatives of u appear in the equation. Then $\mathbf{y}$ is a state vector for

the system for the input set which consists of time functions involving up to the delta function.

State variables in analog computer simulation of differential equation system If an analog computer is used to solve a system of differential equations, the output signals of the integrators constitute a set of state variables because, in an analog computer simulation of an nth-order differential equation system, the output signal of the system is always uniquely determined by the initial conditions (which are set at the n integrators) and the input function (which is a continuous or discontinuous function of time). This means that the n initial conditions which are set at the n integrators at $t = t_0$ are a set of state variables evaluated at $t = t_0$. Hence for $t \geqslant t_0$ the output signals from the n integrators constitute a set of state variables.

Consider, for example, the nth-order differential equation

$$\overset{(n)}{x} + a_1 \overset{(n-1)}{x} + \cdots + a_{n-1}\dot{x} + a_n x = u$$

By properly choosing scale factors, we can make the coefficients of the

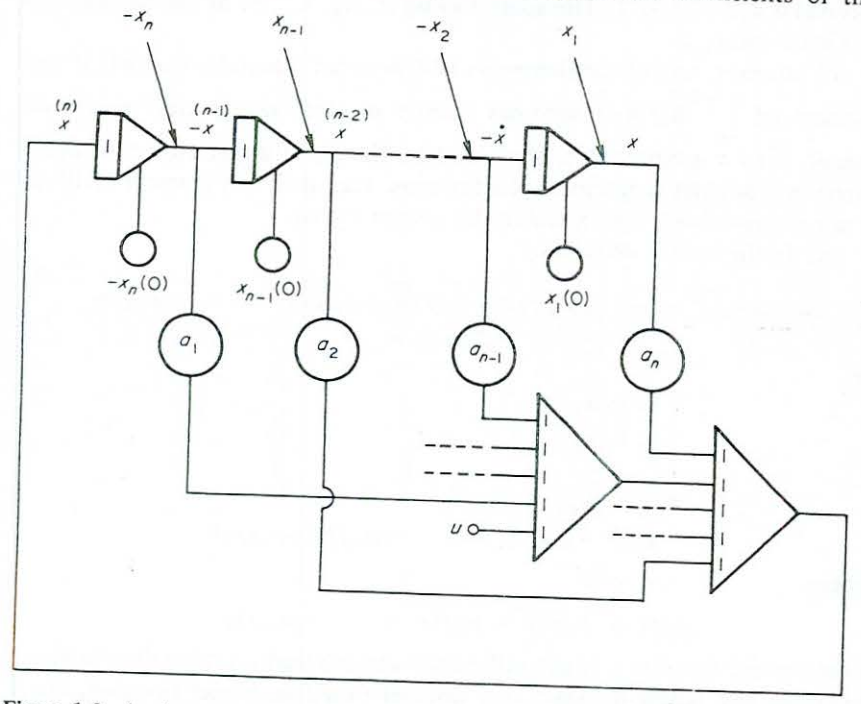


Figure 1-8 Analog computer diagram for nth-order differential equation $\overset{(n)}{x} + a_1 \overset{(n-1)}{x} + \cdots + a_{n-1}\dot{x} + a_n x = u$.

differential equation to be equal to or smaller than unity. Therefore, we can assume that such a change of time scale is already done and the coefficients $a_1, a_2, \ldots, a_n$ are equal to or smaller than unity. Figure 1-8 is a diagram of analog computer simulation for this nth-order differential equation. In this diagram the output signals from the integrators are denoted by $x_1, -x_2, \ldots, -x_n$. Here, $x_1, x_2, \ldots, x_n$ are a set of state variables.

As an another example, consider a nonlinear system defined by

$$\ddot{x} + f(\dot{x}) + x = 0$$

where $f(\dot{x})$ is a nonlinear function of $\dot{x}$. We assume that $f(\dot{x})$ is a function that can be simulated by means of an analog computer. Figure 1-9 shows an analog computer diagram for this system. The output signals from the integrators are denoted by x_1 and $-x_2$. Here, x_1 and x_2 are a set of state variables.

Note that what we have stated here also applies to the differential equation system defined by Eq. (1-17). We shall give a detailed discussion of an analog computer simulation of the system of Eq. (1-17) in Chapter 4.

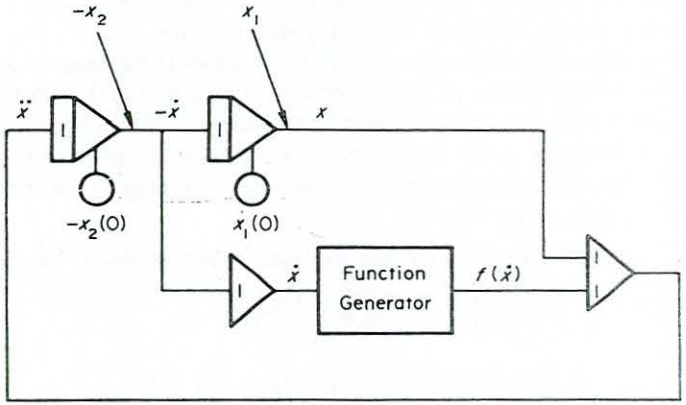


Figure 1-9 Analog computer diagram for system defined by $\ddot{x} + f(\dot{x}) + x = 0$.

Nonuniqueness of set of state variables A set of state variables can be associated with a given system in many ways. In other words there exist a number of different sets of state variables for a given system for a given input set. Which of these is most relevant depends on the individual situation, that is, the nature of the problem, the nature of the input set, etc. If, however, it is desired to construct physically an optimal controller synthesized by optimal control theory, then it is necessary to choose measurable signals as a set of state variables because actual measurement of the state variables is required for realization of an optimal control law.

Suppose that $x_1, x_2, \ldots, x_n$ are a set of state variables. We may take as another set of state variables any set of functions

$$\hat{x}_1 = X_1(x_1, x_2, \ldots, x_n)$$
$$\hat{x}_2 = X_2(x_1, x_2, \ldots, x_n)$$
$$\cdots \quad \cdots \quad \cdots$$
$$\hat{x}_n = X_n(x_1, x_2, \ldots, x_n)$$

provided that for every set of values $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n$ there corresponds a unique set of values $x_1, x_2, \ldots, x_n$, and vice versa.

As just stated, a variety of sets of state variables can be obtained by changing the coordinate system in the state space. Changing the set of state variables will cause changes in the state space equation of the system. Hence the selection of a set of state variables affects greatly the analytic and computational effort required.

Note that choosing a particular set of state variables, though they may not be physically measurable, may simplify greatly the complexity of state space representation of the system. Simplification of the state space equation is discussed in great detail in Chapters 3 and 4.

1-4 OPTIMAL CONTROL SYSTEMS

In classical control theory, attention was focused on the system's absolute stability, transient performance, such as rise time, settling time, maximum overshoot, etc., and steady state accuracy.

Recently, demands for control systems of increasingly high performance have led to increasingly stringent requirements on specifications of transient and steady state responses. Synthesis of control systems which meet stringent requirements on specifications by classical control theory is, in fact, an exceedingly difficult job. In particular, if the performance requirements indicate a need of a time-varying control law, classical control theory is completely inadequate. For such problems, synthesis methods based on optimal control theory offer distinct advantages over classical synthesis methods.

In designing a high-performance system, we strive for an optimal configuration subject to the many and diverse engineering constraints which are imposed. An optimal over-all configuration of the system evolves from a consideration of various factors, such as performance, efficiency, economy, reliability, stability, etc.

This section deals with the basic problem of optimal control systems. *Optimal control* of a plant is finding the present control decision, subject to certain constraints, that will minimize some measure of deviation from

the ideal performance. To make the problem meaningful, we must define a quantitative index of performance precisely. This is one of the most important aspects of the optimal control problem. Note that the problem of defining a performance index is closely related to the problem of deriving a mathematical model of a given plant.

In defining a reasonable performance index, physical intuition plays an important role. Some performance indexes lead to problems which are mathematically tractable; others lead to problems which cannot be solved even by use of digital computers. An example of many performance indexes which are mathematically tractable if used with linear systems is the generalized quadratic performance index, such as the integrated squared error between the actual state and the desired state during a fixed time interval. Constraints may be on the magnitude of the control vector, on the available control energy, or on both.

Once the performance index is defined, the best possible system is that which minimizes this index subject to the constraints imposed on the system. Many investigations in the literature have been concerned with (a) optimal control of linear systems with various quadratic performance indexes, subject to constraints on the control vector and/or the state vector, (b) time optimal control of linear systems, subject to constraints imposed on the control vector.

Problems which have been treated by optimal control theory may be formulated as follows:

1. A plant to be controlled can be adequately represented by a mathematical model of the form

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, t)$$

or its discrete-time equivalence.

2. The desired (or ideal) state of the plant or the desired (or ideal) output of the system is given.
3. A performance index is given. It has the form of

$$J = \int_{t_0}^{T+t_0} L(\mathbf{x}, \mathbf{u}, t)\, dt$$

or its equivalent discrete-time expression. The integrand $L(\mathbf{x}, \mathbf{u}, t)$ in the integral expression is called the *loss function* and represents a measure of the instantaneous deviation from the desired (or ideal) performance.

4. Constraints on the control vector $\mathbf{u}$ and/or the state vector $\mathbf{x}$ of the plant are given.

Once the optimal control problem is mathematically formulated a variety of techniques for the solution is available, including calculus of variations,

dynamic programming, Pontriagin's maximum principle, method of steepest descents, etc. We shall not present various optimal control problems in this book. Instead, we shall treat in Chapter 9 one particular but basic type of optimal control problem; namely, optimal control of linear systems with quadratic or Hermitian performance indexes.

Question of physically constructing optimal controller The essential point is whether we should actually build optimal controllers suggested by optimal control theory. The question arises because implementing the optimal controller presents serious engineering and economic problems.

For linear plants with a mathematically tractable performance index we may be able to obtain analytically the optimal control law. The optimal control law may, however, be extremely complicated, in the sense that it is time-varying and/or nonlinear with extreme complexity, etc. The feasibility of actually building an optimal control system cannot be decided without engineering and economic considerations. (Realization of an optimal controller may be accomplished only by means of extremely complicated special purpose computers.)

Except where optimal performance can pay real dividends, as in military and astronautical applications, we should not build extremely complicated optimal controllers. Systems synthesized by optimal control theory are rather a basis for defining the ultimate performance limitation under the given performance index. In general, we must measure the degree to which the performance of the optimal control system exceeds that of the nonoptimal control system before deciding whether to build the optimal control system or something inferior but simpler.

In many industrial systems, if the difference between the performance of the system with the optimal but complex controller (a complex special purpose computer) and that with a simple and practical (but nonoptimal) controller (a conventional controller or possibly a special purpose computer) is not great, the latter may be preferable.

Note that, besides economic considerations, application of optimal controllers is hampered because of the lack of an effective computational method to use in on-line determination of the optimal control signals. Except in simple cases, computation necessary for determining the present optimal control decision is usually quite involved. The technique of dynamic programming presents a possible solution to this problem, but it is, at present, limited to simple cases.

It must, however, be emphasized that investigations based on optimal control theory will greatly help in understanding and developing theories which eventually will lead to practical systems close to optimal. Such systems may possibly include computers for the generation of control signals. Design of such systems cannot possibly be accomplished by classical control theory.

1-5 CLASSICAL CONTROL THEORY VERSUS MODERN CONTROL THEORY

In concluding this chapter we shall compare classical control theory and modern control theory from various viewpoints.

1. Classical control theory is best suited for design of single-input single-output linear time-invariant systems. The design is based on the transient response and frequency response characteristics of the system. In classical control theory severe limitations and difficulties exist for design of multivariable control systems and time-varying control systems. Design by modern control theory based on state space concepts can be carried out for a class of inputs instead of the impulse function, step function, or sinusoidal function. Modern control theory can be applied to the design of linear multivariable control systems and linear time-varying control systems that are optimal with given performance indexes.

2. Experimentally, a complex system can be characterized easily in terms of frequency response curves. Frequency response tests are, in general, simple and can be made accurate by use of accurate sinusoidal signal generators and precise sinusoidal measurement equipments. It is usually difficult to write a set of differential equations representing the dynamics of a complex system so that modern control theory can be applied directly. Hence, from this viewpoint, classical control theory is more convenient than modern control theory.

3. In classical control theory, algebraic manipulation of the parameters can be made before numerical values are substituted. This is not the case in modern control theory because of its extensive use of vector matrix notation. Vector matrix notation greatly simplifies description of system equations. Notational and conceptual simplicity of vector matrix operations is essential for theoretical work. Although the analytical solution in vector matrix notation may seem simple, if the numerical solution is required, computation becomes extremely involved. Except in simple cases, we must resort to digital computers for necessary computations. (In obtaining computer solutions we must give the numerical values for all parameters.)

4. Inclusion of the initial conditions in the system design is impossible in classical control theory. Modern control theory, on the other hand, enables us to include the system's initial conditions in the design.

5. For high-frequency noise rejection in control systems, classical frequency domain design techniques are more convenient than state space techniques.

From the brief comparisons just made, we can conclude that although modern control theory is more general than classical control theory, the latter will not be discarded unless the former becomes, in future, better in every respect of control system analysis and synthesis.

Outline of text In this book we shall be concerned only with deterministic systems. The system is called *deterministic* if the future states are completely determined by the present state provided that the future inputs are known.

The contents of the text are as follows: Chapter 1 presents an introduction to state space analysis of control systems. Chapter 2 reviews introductory material on matrices and vectors. Chapter 3 discusses reduction of square matrices to the Jordan canonical form and coordinate transformation in state spaces. Chapter 4 deals with state space representation of control systems. Chapter 5 presents additional materials on matrix analysis which are needed for studying Chapters 6, 7, 8, and 9. Chapter 6 gives formal solutions of state space equations. Chapter 7 presents detailed discussions of controllability and observability of control systems. Chapter 8 treats stability analyses based on the second method of Liapunov and on Routh-Hurwitz stability criteria. Some applications of the second method of Liapunov to design problems are included. Finally, Chapter 9 discusses analytical design of optimal control systems based on quadratic performance indexes.

Throughout the book an attempt has been made to present basic mathematical techniques necessary for understanding modern control theory in the simplest and easiest way possible. A number of simple examples are provided at the end of each chapter as Exercise Problems and Solutions. Those who are as yet unfamiliar with modern control theory will find these examples useful for understanding the materials presented in the book and for learning techniques for analyzing problems in modern control theory.

EXERCISE PROBLEMS AND SOLUTIONS

Problem A-1-1 Hamilton showed that the first variation of the time integral of the difference between the kinetic energy T and the potential energy V of a dynamic system is zero, or

$$\delta \int_{t_1}^{t_2} (T - V)\, dt = 0$$

This is known as *Hamilton's principle*. Derive Hamilton's principle from d'Alembert's principle, which states that the applied force, together with the forces of inertia, forms a system in equilibrium. For a system consisting of n particles, d'Alembert's principle can be stated analytically as

$$\sum_{i=1}^{n} \left[\left(m_i \frac{d^2 x_i}{dt^2} - X_i \right) \delta x_i + \left(m_i \frac{d^2 y_i}{dt^2} - Y_i \right) \delta y_i + \left(m_i \frac{d^2 z_i}{dt^2} - Z_i \right) \delta z_i \right] = 0 \tag{1-22}$$

Solution Since

$$\begin{aligned} \frac{d^2 x_i}{dt^2} \delta x_i &= \frac{d}{dt}\left(\frac{dx_i}{dt} \delta x_i \right) - \frac{dx_i}{dt} \frac{d\, \delta x_i}{dt} \\ &= \frac{d}{dt}\left(\frac{dx_i}{dt} \delta x_i \right) - \frac{1}{2} \delta \left(\frac{dx_i}{dt} \right)^2 \end{aligned}$$

Eq. (1-22) becomes

$$\begin{aligned} \frac{d}{dt} \sum_{i=1}^{n} \left[m_i \left(\frac{dx_i}{dt} \delta x_i + \frac{dy_i}{dt} \delta y_i + \frac{dz_i}{dt} \delta z_i \right) \right] \\ = \frac{1}{2} \delta \sum_{i=1}^{n} \left\{ m_i \left[\left(\frac{dx_i}{dt} \right)^2 + \left(\frac{dy_i}{dt} \right)^2 + \left(\frac{dz_i}{dt} \right)^2 \right] \right\} \\ + \sum_{i=1}^{n} (X_i\, \delta x_i + Y_i\, \delta y_i + Z_i\, \delta z_i) \end{aligned} \tag{1-23}$$

Here,

$$\frac{1}{2} \sum_{i=1}^{n} m_i \left[\left(\frac{dx_i}{dt} \right)^2 + \left(\frac{dy_i}{dt} \right)^2 + \left(\frac{dz_i}{dt} \right)^2 \right]$$

is the kinetic energy T of the system. If there is a potential function we have

$$\sum_{i=1}^{n} (X_i\, \delta x_i + Y_i\, \delta y_i + Z_i\, \delta z_i) = -\delta V$$

Integrating Eq. (1-23) with respect to t from t_1 to t_2,

$$\sum_{i=1}^{n} \left[m_i \left(\frac{dx_i}{dt} \delta x_i + \frac{dy_i}{dt} \delta y_i + \frac{dz_i}{dt} \delta z_i \right) \right]_{t_1}^{t_2} = \int_{t_1}^{t_2} (\delta T - \delta V)\, dt$$

If the positions of the particles of the system are fixed at $t = t_1$ and $t = t_2$, the variations δx_i, δy_i, δz_i vanish for $t = t_1$ and $t = t_2$. The left-hand side of the preceding equation then vanishes. Hence

$$\delta \int_{t_1}^{t_2} (T - V)\, dt = 0$$

This is Hamilton's principle.

Problem A-1-2 Suppose that there are r masses whose coordinates are (x_1, x_2, x_3), (x_4, x_5, x_6), . . . , $(x_{3r-2}, x_{3r-1}, x_{3r})$. Suppose also that forces acting on r masses are (X_1, X_2, X_3), (X_4, X_5, X_6), . . . , $(X_{3r-2}, X_{3r-1}, X_{3r})$. Newton's equations of motion are given by

$$m_i \ddot{x}_i = X_i \qquad (i = 1, 2, \ldots, 3r)$$

where $m_1 = m_2 = m_3$; . . . ; $m_{3r-2} = m_{3r-1} = m_{3r}$. If there are s (where $s < 3r$) constraint equations of the form

$$x_i = x_i(q_1, q_2, \ldots, q_n)$$

where $n = 3r - s$, then show that the kinetic energy of the system can be a function of both the $\dot{q}_i$ and q_i $(i = 1, 2, \ldots, n)$.

Solution Notice that

$$\dot{x}_i = \sum_{j=1}^{n} \frac{\partial x_i}{\partial q_j} \dot{q}_j$$

Hence,

$$\begin{aligned} T = \sum_{i=1}^{3r} \frac{m_i}{2} \dot{x}_i^2 &= \sum_{i=1}^{3r} \frac{m_i}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} \frac{\partial x_i}{\partial q_j} \frac{\partial x_i}{\partial q_k} \dot{q}_j \dot{q}_k \\ &= \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} \left(\sum_{i=1}^{3r} m_i \frac{\partial x_i}{\partial q_j} \frac{\partial x_i}{\partial q_k} \right) \dot{q}_j \dot{q}_k \\ &= \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} Q_{jk} \dot{q}_j \dot{q}_k \end{aligned}$$

where

$$Q_{jk} = \sum_{i=1}^{3r} m_i \frac{\partial x_i}{\partial q_j} \frac{\partial x_i}{\partial q_k}$$

Note that the Q_{jk} are constants or functions of q_i $(i = 1, 2, \ldots, n)$. Hence T can be a function of the $\dot{q}_i$ and q_i.

Problem A-1-3 Derive Lagrange's equations of motion

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = 0 \qquad (i = 1, 2, \ldots, n)$$

from Hamilton's principle

$$\delta \int_{t_1}^{t_2} L \, dt = 0$$

where $L = T - V =$ Lagrangian

Solution The kinetic energy T is a quadratic function of $\dot{q}_1, \dot{q}_2, \ldots, \dot{q}_n$ and can be written as

$$T = \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} Q_{jk} \dot{q}_j \dot{q}_k$$

where the Q_{jk} are constants or functions of q_i $(i = 1, 2, \ldots, n)$. The potential energy V is a function only of q_i and not a function of $\dot{q}_i$. From Hamilton's principle, we obtain

$$\int_{t_1}^{t_2} \left\{ \sum_{i=1}^{n} \left[\frac{\partial (T - V)}{\partial q_i} \delta q_i + \frac{\partial (T - V)}{\partial \dot{q}_i} \delta \dot{q}_i \right] \right\} dt = 0$$

By integrating the second term by parts, noting that

$$\delta \dot{q}_i = \delta \frac{dq_i}{dt} = \frac{d}{dt} \delta q_i$$

and that the δq_i vanish at $t = t_1$ and $t = t_2$, we obtain

$$\int_{t_1}^{t_2} \left\{ \sum_{i=1}^{n} \left[\frac{\partial (T - V)}{\partial q_i} - \frac{d}{dt} \left(\frac{\partial (T - V)}{\partial \dot{q}_i} \right) \right] \delta q_i \right\} dt = 0$$

Since the preceding equation must hold for any δq_i, the coefficient of every δq_i must vanish, or

$$\frac{\partial (T - V)}{\partial q_i} - \frac{d}{dt} \left(\frac{\partial (T - V)}{\partial \dot{q}_i} \right) = 0 \qquad (i = 1, 2, \ldots, n)$$

If we substitute L for $T - V$ in the foregoing equation

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = 0 \qquad (i = 1, 2, \ldots, n)$$

These are Lagrange's equations of motion.

Problem A-1-4 Derive Hamilton's equations of motion

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \qquad \dot{p}_i = -\frac{\partial H}{\partial q_i} \qquad (i = 1, 2, \ldots, n)$$

where $H = T + V =$ Hamiltonian

from Hamilton's principle

$$\delta \int_{t_1}^{t_2} L \, dt = 0$$

where $L = T - V =$ Lagrangian

The variables p_i are defined by

$$p_i = \frac{\partial T}{\partial \dot{q}_i} \qquad (i = 1, 2, \ldots, n)$$

Solution Since the Hamiltonian function H is defined by

$$H = T + V$$

the Lagrangian function can be written as

$$L = T - V = 2T - H$$

Here T and H are functions of q_i and p_i. Since T can be written as

$$T = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \dot{q}_i \dot{q}_j \qquad (a_{ij} = a_{ji})$$

where the a_{ij} are constants or are functions of $q_1, q_2, \ldots, q_n$, we obtain

$$\begin{aligned} p_1 &= \frac{\partial T}{\partial \dot{q}_1} = a_{11}\dot{q}_1 + a_{12}\dot{q}_2 + \cdots + a_{1n}\dot{q}_n \\ p_2 &= \frac{\partial T}{\partial \dot{q}_2} = a_{21}\dot{q}_1 + a_{22}\dot{q}_2 + \cdots + a_{2n}\dot{q}_n \\ &\cdots \qquad \cdots \qquad \cdots \\ p_n &= \frac{\partial T}{\partial \dot{q}_n} = a_{n1}\dot{q}_1 + a_{n2}\dot{q}_2 + \cdots + a_{nn}\dot{q}_n \end{aligned}$$

We see that

$$2T = \sum_{i=1}^{n} \frac{\partial T}{\partial \dot{q}_i} \dot{q}_i$$

Hamilton's principle then becomes

$$\delta\int_{t_1}^{t_2}(2T-H)\,dt=\delta\int_{t_1}^{t_2}\left(\sum_{i=1}^{n}\frac{\partial T}{\partial\dot{q}_i}\dot{q}_i-H\right)dt$$

$$=\delta\int_{t_1}^{t_2}\left(\sum_{i=1}^{n}p_i\dot{q}_i-H\right)dt$$

$$=\int_{t_1}^{t_2}\sum_{i=1}^{n}\left(\dot{q}_i\,\delta p_i+p_i\,\delta\dot{q}_i-\frac{\partial H}{\partial p_i}\delta p_i-\frac{\partial H}{\partial q_i}\delta q_i\right)dt=0$$

Noting that

$$\int_{t_1}^{t_2}p_i\,\delta\dot{q}_i\,dt=p_i\,\delta q_i\Big|_{t=t_1}^{t=t_2}-\int_{t_1}^{t_2}\dot{p}_i\,\delta q_i\,dt$$

and δq_i vanish at $t=t_1$ and $t=t_2$, we obtain

$$\int_{t_1}^{t_2}\sum_{i=1}^{n}\left[\left(\dot{q}_i-\frac{\partial H}{\partial p_i}\right)\delta p_i-\left(\dot{p}_i+\frac{\partial H}{\partial q_i}\right)\delta q_i\right]dt=0 \qquad (1\text{–}24)$$

Since V does not involve p_i

$$\frac{\partial H}{\partial p_i}=\frac{\partial T}{\partial p_i}$$

It can be shown that

$$\frac{\partial T}{\partial p_i}=\dot{q}_i$$

We will demonstrate this for a simple case where

$$T=\tfrac{1}{2}(a_{11}\dot{q}_1^2+2a_{12}\dot{q}_1\dot{q}_2+a_{22}\dot{q}_2^2)$$

By definition, p_1 and p_2 are obtained as

$$p_1=\frac{\partial T}{\partial\dot{q}_1}=a_{11}\dot{q}_1+a_{12}\dot{q}_2$$

$$p_2=\frac{\partial T}{\partial\dot{q}_2}=a_{12}\dot{q}_1+a_{22}\dot{q}_2$$

Solving for $\dot{q}_1$ and $\dot{q}_2$, we obtain

$$\dot{q}_1=\frac{a_{22}p_1-a_{12}p_2}{a_{11}a_{22}-a_{12}^2},\qquad \dot{q}_2=\frac{-a_{12}p_1+a_{11}p_2}{a_{11}a_{22}-a_{12}^2}$$

Note that $a_{11}a_{22}-a_{12}^2>0$, since T is a positive definite quantity. (See p. 79.) The kinetic energy then can be written in terms of p_1 and p_2 as follows:

$$T=\frac{1}{2(a_{11}a_{22}-a_{12}^2)}(a_{22}p_1^2-2a_{12}p_1p_2+a_{11}p_2^2)$$

By differentiating T with respect to p_1 and p_2, respectively, we obtain

$$\frac{\partial T}{\partial p_1}=\frac{a_{22}p_1-a_{12}p_2}{a_{11}a_{22}-a_{12}^2}=\dot{q}_1$$

$$\frac{\partial T}{\partial p_2}=\frac{-a_{12}p_1+a_{11}p_2}{a_{11}a_{22}-a_{12}^2}=\dot{q}_2$$

or

$$\frac{\partial T}{\partial p_i}=\dot{q}_i \qquad (i=1,2)$$

We can easily extend this result for $i=1,2,\ldots,n$. It is simpler, however, to use matrix algebra to show this for a general case where $i=1,2,\ldots,n$. (See Prob. A-5-13). We thus obtain

$$\dot{q}_i=\frac{\partial T}{\partial p_i}=\frac{\partial H}{\partial p_i}$$

The coefficients of the δp_i in Eq. (1–24) vanish. Hence Eq. (1–24) becomes

$$\int_{t_1}^{t_2}\left[\sum_{i=1}^{n}\left(\dot{p}_i+\frac{\partial H}{\partial q_i}\right)\delta q_i\right]dt=0$$

Since the δq_i are arbitrary, their coefficients must be equal to zero, or

$$\dot{p}_i+\frac{\partial H}{\partial q_i}=0 \qquad (i=1,2,\ldots,n)$$

We have thus derived Hamilton's equations of motion:

$$\dot{q}_i=\frac{\partial H}{\partial p_i},\qquad \dot{p}_i=-\frac{\partial H}{\partial q_i} \qquad (i=1,2,\ldots,n)$$

Problem A-1-5 Figure 1–10 shows a double pendulum. Obtain the Hamiltonian function H for this systems. Assume that the pendulum swings in one plane.

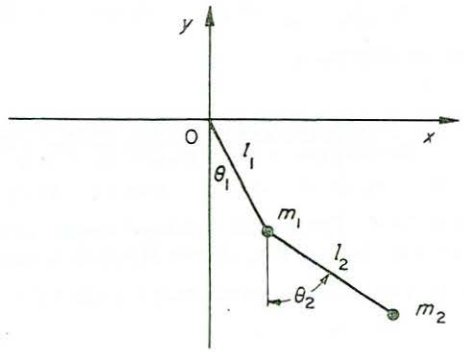


Figure 1-10 Double pendulum.

Solution Let us define generalized coordinates q_1 and q_2 by

$$q_1=\theta_1,\qquad q_2=\theta_2$$

In this system the mass m_1 has risen from its equilibrium position a distance $l_1(1-\cos q_1)$ and the mass m_2 has risen from its equilibrium position a distance $l_1(1-\cos q_1)+l_2(1-\cos q_2)$. The potential energy is therefore given by

$$V=m_1gl_1(1-\cos q_1)+m_2g[l_1(1-\cos q_1)+l_2(1-\cos q_2)]$$

Here we have regarded $q_1=0$ and $q_2=0$ as the zero potential energy datum. The x and y coordinates of the displacement (x_1, y_1) of the mass m_1 are given by

$$x_1=l_1\sin q_1,\qquad y_1=-l_1\cos q_1$$

and those of the mass m_2 are given by

$$x_2=l_1\sin q_1+l_2\sin q_2,\qquad y_2=-l_1\cos q_1-l_2\cos q_2$$

Hence, the square of the velocity of the mass m_1 is given by

$$v_1^2 = \dot{x}_1^2 + \dot{y}_1^2 = (l_1\dot{q}_1)^2$$

and the square of the velocity of the mass m_2 is given by

$$\begin{aligned} v_2^2 &= \dot{x}_2^2 + \dot{y}_2^2 \\ &= (l_1\dot{q}_1)^2 + (l_2\dot{q}_2)^2 + 2l_1l_2\dot{q}_1\dot{q}_2\cos(q_1 - q_2) \end{aligned}$$

Therefore, the kinetic energy is given by

$$\begin{aligned} T &= \left(\frac{m_1}{2}l_1^2 + \frac{m_2}{2}l_1^2\right)\dot{q}_1^2 + m_2l_1l_2\dot{q}_1\dot{q}_2\cos(q_1 - q_2) + \left(\frac{m_2}{2}l_2^2\right)\dot{q}_2^2 \\ &= \frac{1}{2}(a_{11}\dot{q}_1^2 + 2a_{12}\dot{q}_1\dot{q}_2 + a_{22}\dot{q}_2^2) \end{aligned}$$

where

$$\begin{aligned} a_{11} &= (m_1 + m_2)l_1^2 \\ a_{12} &= m_2l_1l_2\cos(q_1 - q_2) \\ a_{22} &= m_2l_2^2 \end{aligned}$$

The generalized momenta are

$$\begin{aligned} p_1 &= \frac{\partial T}{\partial \dot{q}_1} = a_{11}\dot{q}_1 + a_{12}\dot{q}_2 \\ p_2 &= \frac{\partial T}{\partial \dot{q}_2} = a_{12}\dot{q}_1 + a_{22}\dot{q}_2 \end{aligned}$$

From Prob. A-1-4, T can be written in terms of p_1 and p_2 as follows:

$$T = \frac{1}{2(a_{11}a_{22} - a_{12}^2)}(a_{22}p_1^2 - 2a_{12}p_1p_2 + a_{11}p_2^2)$$

Hence the Hamiltonian function is

$$\begin{aligned} H &= T + V \\ &= \frac{m_2l_2^2p_1^2 - 2m_2l_1l_2p_1p_2\cos(q_1 - q_2) + (m_1 + m_2)l_1^2p_2^2}{2(m_2l_1^2l_2^2)[m_1 + m_2 - m_2\cos^2(q_1 - q_2)]} \\ &\quad + (m_1 + m_2)gl_1(1 - \cos q_1) + m_2gl_2(1 - \cos q_2) \end{aligned}$$

Problem A-1-6 Obtain the potential energy and kinetic energy of the system shown in Fig. 1–11. Then, derive Hamilton's equations of motion.

Solution Let us define generalized coordinates q_1 and q_2 as follows:

$$q_1 = x_1, \qquad q_2 = x_2$$

The kinetic energy of the system is

$$T = \tfrac{1}{2}m_1\dot{q}_1^2 + \tfrac{1}{2}m_2\dot{q}_2^2$$

The potential energy of the system is

$$V = \tfrac{1}{2}k_1q_1^2 + \tfrac{1}{2}k_2(q_2 - q_1)^2$$

(Note that here the generalized coordinates q_1 and q_2 are taken from the static equilibrium position.) The generalized momenta are defined by

$$\begin{aligned} p_1 &= \frac{\partial T}{\partial \dot{q}_1} = m_1\dot{q}_1 \\ p_2 &= \frac{\partial T}{\partial \dot{q}_2} = m_2\dot{q}_2 \end{aligned}$$

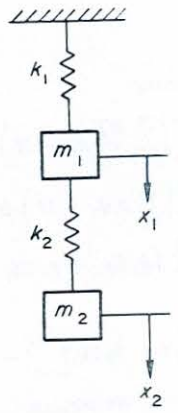


Figure 1-11 Mechanical vibratory system.

Thus

$$\dot{q}_1 = \frac{p_1}{m_1}, \qquad \dot{q}_2 = \frac{p_2}{m_2}$$

$$T = \frac{1}{2}m_1\left(\frac{p_1}{m_1}\right)^2 + \frac{1}{2}m_2\left(\frac{p_2}{m_2}\right)^2$$

The Hamiltonian function for this system is

$$H = \frac{p_1^2}{2m_1} + \frac{p_2^2}{2m_2} + \frac{1}{2}k_1q_1^2 + \frac{1}{2}k_2(q_2 - q_1)^2$$

Hamilton's equations of motion are

$$\begin{aligned} \dot{q}_1 &= \frac{\partial H}{\partial p_1} = \frac{p_1}{m_1} \\ \dot{q}_2 &= \frac{\partial H}{\partial p_2} = \frac{p_2}{m_2} \\ \dot{p}_1 &= -\frac{\partial H}{\partial q_1} = -k_1q_1 + k_2(q_2 - q_1) \\ \dot{p}_2 &= -\frac{\partial H}{\partial q_2} = -k_2(q_2 - q_1) \end{aligned}$$

which are equivalent to the following familiar equations

$$\begin{aligned} m_1\ddot{x}_1 + k_1x_1 + k_2(x_1 - x_2) &= 0 \\ m_2\ddot{x}_2 + k_2(x_2 - x_1) &= 0 \end{aligned}$$

Problem A-1-7 Consider the mechanical vibratory system shown in Fig. 1–12. By defining a set of state variables, obtain a state space equation of the system.

Solution The equations of motion are given by

$$\begin{aligned} m\ddot{y}_1 + ky_1 - k(y_2 - y_1) &= 0 \\ m\ddot{y}_2 + k(y_2 - y_1) + ky_2 &= 0 \end{aligned}$$

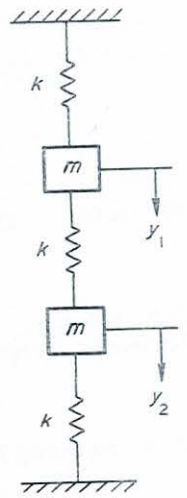


Figure 1-12 Mechanical vibratory system.

By defining

$$x_1 = y_1, \qquad x_2 = \dot{y}_1, \qquad x_3 = y_2, \qquad x_4 = \dot{y}_2$$

we obtain

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = -\frac{2k}{m}x_1 + \frac{k}{m}x_3$$
$$\dot{x}_3 = x_4$$
$$\dot{x}_4 = \frac{k}{m}x_1 - \frac{2k}{m}x_3$$

In this system x_1, x_2, x_3, and x_4 are a set of state variables. Using vector matrix notation we obtain

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\frac{2k}{m} & 0 & \frac{k}{m} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{k}{m} & 0 & -\frac{2k}{m} & 0 \end{bmatrix}$$

Problem A-1-8 Obtain a state space equation of the system shown in Fig. 1–13.

Solution The equation relating x and u is

$$\ddot{x} + x + a(t)x + x^2 = u$$

Define $x_1 = x$, and $x_2 = \dot{x}$. Then, a state space representation of the system becomes

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = -x_1 - a(t)x_1 - x_1^2 + u$$

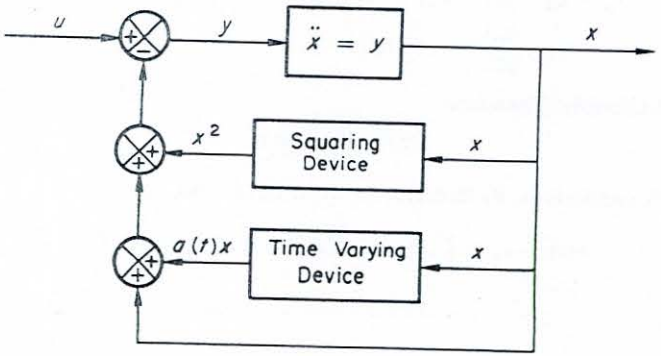


Figure 1-13 Nonlinear system.

Problem A-1-9 (1–3)[1] Derive the equation of the three axes reaction-wheel system shown in Fig. 1–14. Assume that the reaction-wheel spin axes are aligned with principal inertia axes and that the effect of small misalignments between these two sets of axes can be neglected. Assume also that the inertial ellipsoid of the spacecraft is time invariant.

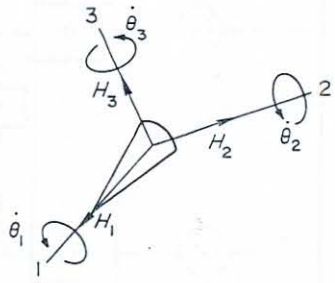


Figure 1-14 Three axes reaction-wheel system.

In the figure, the three axes shown are rigid spacecraft axes. In order to derive the equation, define

$\dot{\theta}_i$ = angular velocity about the ith axis (that is, the quantity measured by a rate gyro about the ith axis)
T_i = external torque about the ith axis
H_i = total angular momentum component along the ith axis
$= h_i + I_i\dot{\theta}_i + g_i$

where h_i = angular momentum of reaction wheel about the ith axis
$I_i\dot{\theta}_i$ = angular momentum of spacecraft body about the ith axis
g_i = ith angular momentum component of rotating parts within spacecraft (such as rate gyros) excluding h_i

Solution

$$\begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix} = \begin{bmatrix} \dot{H}_1 \\ \dot{H}_2 \\ \dot{H}_3 \end{bmatrix} + \begin{bmatrix} H_3\dot{\theta}_2 - H_2\dot{\theta}_3 \\ H_1\dot{\theta}_3 - H_3\dot{\theta}_1 \\ H_2\dot{\theta}_1 - H_1\dot{\theta}_2 \end{bmatrix}$$
$$= \begin{bmatrix} \dot{h}_1 \\ \dot{h}_2 \\ \dot{h}_3 \end{bmatrix} + \begin{bmatrix} I_1\ddot{\theta}_1 \\ I_2\ddot{\theta}_2 \\ I_3\ddot{\theta}_3 \end{bmatrix} + \begin{bmatrix} \dot{g}_1 \\ \dot{g}_2 \\ \dot{g}_3 \end{bmatrix}$$
$$+ \begin{bmatrix} (h_3 + g_3)\dot{\theta}_2 - (h_2 + g_2)\dot{\theta}_3 \\ (h_1 + g_1)\dot{\theta}_3 - (h_3 + g_3)\dot{\theta}_1 \\ (h_2 + g_2)\dot{\theta}_1 - (h_1 + g_1)\dot{\theta}_2 \end{bmatrix} + \begin{bmatrix} (I_3 - I_2)\dot{\theta}_2\dot{\theta}_3 \\ (I_1 - I_3)\dot{\theta}_1\dot{\theta}_3 \\ (I_2 - I_1)\dot{\theta}_1\dot{\theta}_2 \end{bmatrix}$$

Problem A-1-10 Figure 1–15 shows a multiple-input multiple-output system. Obtain a state space equation of this system.

Solution

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -11 & 3 \\ 4 & -7 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 10 & -5 \\ -5 & 5 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

Problem A-1-11 Consider the ordinary differential equation of the first order

$$\dot{x}(t) = f(x(t), t) \tag{1–25}$$

[1]See list of references at the end of this chapter.

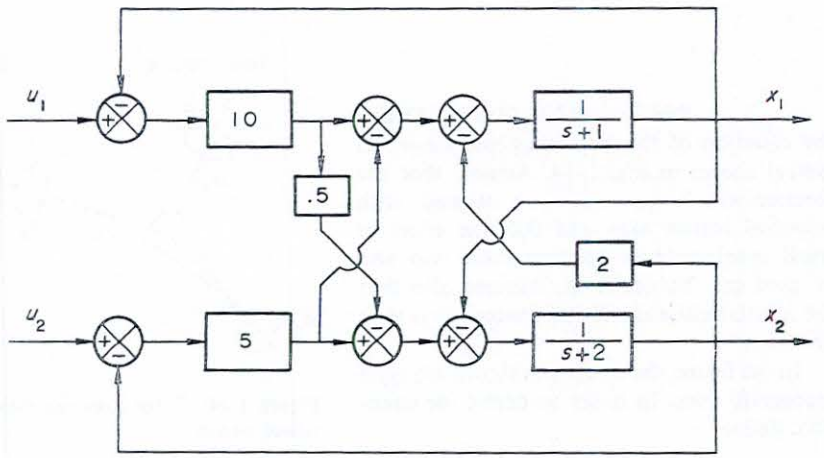


Figure 1-15 Multiple-input multiple-output system.

where $x(t_0) = x_0$

Prove that if $f(x(t), t)$ is single-valued, continuous, and bounded in the region Γ defined by $|x - x_0| < h$, $0 \leqslant t - t_0 < \tau$ and if the following Lipschitz condition:

$$|f(x, t) - f(z, t)| \leqslant K|x - z|$$

where points (x, t) and (z, t) are arbitrary points in the region Γ and K is a positive constant, is satisfied, then the given differential equation has a solution satisfying the initial condition $x(t_0) = x_0$. Prove also that the solution $x(t)$ defined in the interval

$$0 \leqslant t - t_0 < T, \qquad T = \min\left(\tau, \frac{h}{M}\right)$$

where M is the maximum value of $|f(x, t)|$, and satisfying the initial condition is unique.

Solution We shall first prove the existence of the solution. Define $x_0(t), x_1(t), \ldots, x_n(t), \ldots$ in the interval $0 \leqslant t - t_0 < T$ as follows:

(i) $$x_0(t) = x_0$$

(ii) $$x_{n+1}(t) = x_0 + \int_{t_0}^{t} f(x_n(t), t)\, dt \tag{1-26}$$

We must prove that such definition is possible. In order for x_{n+1} to be defined by Eq. (1–26) it is necessary to show that $f(x_n(t), t)$ can be defined in the interval $0 \leqslant t - t_0 < T$ and is continuous. To show this we need prove only that

$$|x_n(t) - x_0| < h, \qquad 0 \leqslant t - t_0 < T \tag{1-27}$$

For $n = 0$, inequality (1–27) is clearly satisfied. Suppose that $|x_{n-1}(t) - x_0| < h$. Then by assumption

$$|f(x_{n-1}(t), t)| \leqslant M$$

Then from Eq. (1–26)

$$|x_n(t) - x_0| = \left|\int_{t_0}^{t} f(x_{n-1}(t), t)\, dt\right| \leqslant M(t - t_0)$$

On the other hand

$$0 \leqslant t - t_0 < T \leqslant \frac{h}{M}$$

Hence we have proved that

$$|x_n(t) - x_0| \leqslant M(t - t_0) < h$$

Thus we have shown that the function series $x_0(t), x_1(t), \ldots, x_n(t), \ldots$ can be defined as given by Eq. (1–26). Note that each $x_i(t)$ is continuous.

We shall next prove that the function series converges to a function as n approaches infinity. To prove this we need to show that

$$|x_{n+1}(t) - x_n(t)| \leqslant \frac{K^n N(t - t_0)^n}{n!} \qquad (n = 0, 1, 2, \ldots) \tag{1-28}$$

where N is a positive number. We shall prove this by mathematical induction. Inequality (1–28) clearly holds for $n = 0$, since x_0 and x_1 are continuous there exists a positive number N such that $|x_1(t) - x_0| = N$. Suppose that inequality (1–28) is true for $n - 1$, then

$$\begin{aligned}
|x_{n+1}(t) - x_n(t)| &= \left|\left(x_0 + \int_{t_0}^{t} f(x_n(t), t)\, dt\right) - \left(x_0 + \int_{t_0}^{t} f(x_{n-1}(t), t)\, dt\right)\right| \\
&= \left|\int_{t_0}^{t} \left[f(x_n(t), t) - f(x_{n-1}(t), t)\right] dt\right| \\
&\leqslant \left|\int_{t_0}^{t} K|x_n(t) - x_{n-1}(t)|\, dt\right| \\
&\leqslant \left|\int_{t_0}^{t} K\frac{K^{n-1}N(t - t_0)^{n-1}}{(n-1)!}\, dt\right| \\
&= K^n N\left|\int_{t_0}^{t} \frac{(t - t_0)^{n-1}}{(n-1)!}\, dt\right| \\
&= \frac{K^n N(t - t_0)^n}{n!}
\end{aligned}$$

Thus, we have proved inequality (1–28). Hence,

$$\begin{aligned}
x_n &= x_0 + (x_1 - x_0) + (x_2 - x_1) + \cdots + (x_n - x_{n-1}) \\
&= x_0 + \sum_{k=0}^{n-1} (x_{k+1} - x_k)
\end{aligned}$$

converges uniformly. Therefore

$$x(t) = \lim_{n \to \infty} x_n(t)$$

exists and is continuous. By letting $n \longrightarrow \infty$ in Eq. (1–26),

$$x(t) = x_0 + \int_{t_0}^{t} f(x(t), t)\, dt, \qquad 0 \leqslant t - t_0 < T \tag{1-29}$$

For $0 \leqslant t - t_0 < T$, we can differentiate Eq. (1–29) with respect to t, or $\dot{x}(t) = f(x(t), t)$. Clearly, $x(t_0) = x_0$. This proves the existence of the solution satisfying the initial condition.

We shall next prove uniqueness of the solution. Suppose that $\hat{x}(t)$ is another solution of Eq. (1–25). Then

$$\hat{x}(t) = x_0 + \int_{t_0}^{t} f(\hat{x}(t), t)\, dt$$

Since $x(t) - \hat{x}(t)$ is continuous in the interval $0 \leqslant t - t_0 < T$, it has a maximum, G. Then we can prove that the following inequality holds true:

$$|x(t) - \hat{x}(t)| \leqslant K^n G \frac{(t - t_0)^n}{n!} \qquad (n = 0, 1, 2, \ldots) \tag{1–30}$$

To prove this, we shall use mathematical induction. Clearly this holds true for $n = 0$. Suppose that it is true for $n - 1$. Then

$$\begin{aligned} |x(t) - \hat{x}(t)| &= \left| \int_{t_0}^{t} \left[f(x(t), t) - f(\hat{x}(t), t) \right] dt \right| \\ &\leqslant \left| \int_{t_0}^{t} K |x(t) - \hat{x}(t)|\, dt \right| \\ &\leqslant \left| \int_{t_0}^{t} K \frac{K^{n-1} G (t - t_0)^{n-1}}{(n-1)!}\, dt \right| \\ &= K^n G \frac{(t - t_0)^n}{n!} \end{aligned}$$

This completes the proof of inequality (1–30). By letting $n \longrightarrow \infty$, $|x(t) - \hat{x}(t)| \longrightarrow 0$. Hence

$$x(t) = \hat{x}(t)$$

This proves uniqueness of the solution.

Problem A-1-12 Consider the following simultaneous first-order differential equations:

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, x_2, t) \\ \dot{x}_2 &= f_2(x_1, x_2, t) \end{aligned} \tag{1–31}$$

where $$x_1(t_0) = x_{10}, \qquad x_2(t_0) = x_{20}$$

Assume that $f_1(x_1, x_2, t)$ and $f_2(x_1, x_2, t)$ are single-valued and continuous in the region Γ defined by

$$|x_1 - x_{10}| < h, \qquad |x_2 - x_{20}| < h, \qquad 0 \leqslant t - t_0 < \tau$$

and $|f_1| \leqslant M$, $|f_2| \leqslant M$ in the region Γ. Assume also that f_1 and f_2 satisfy the following Lipschitz conditions:

$$\begin{aligned} |f_1(x_1, x_2, t) - f_1(z_1, z_2, t)| &\leqslant K(|x_1 - z_1| + |x_2 - z_2|) \\ |f_2(x_1, x_2, t) - f_2(z_1, z_2, t)| &\leqslant K(|x_1 - z_1| + |x_2 - z_2|) \end{aligned}$$

where points (x_1, x_2, t) and (z_1, z_2, t) are arbitrary points in the region Γ and K is a positive constant.

Prove that there exists a unique solution of the differential equation system of

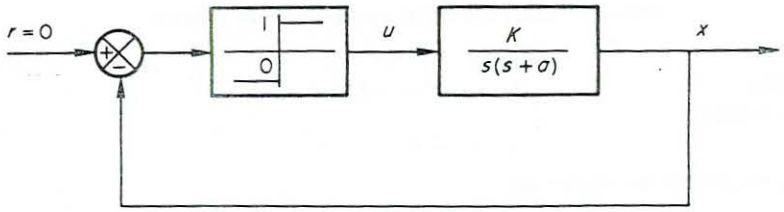
Eq. (1–31) satisfying the given initial condition and defined in the interval

$$0 \leqslant t - t_0 < T, \qquad T = \min\left(\tau, \frac{h}{M}\right)$$

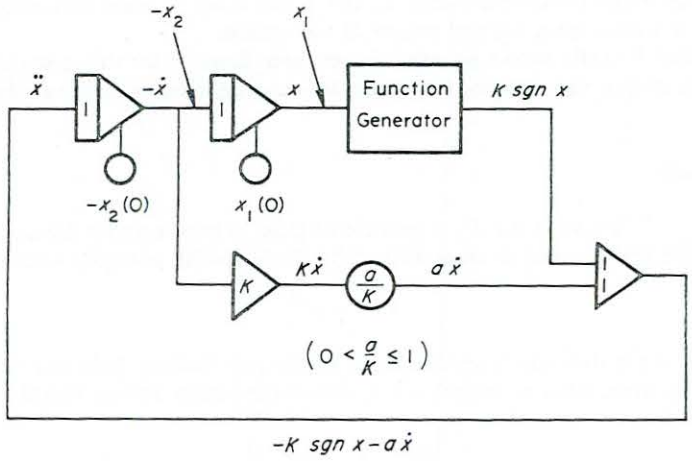
Solution The proof essentially parallels that for Prob. A-1-11. Therefore, we shall give only the outline of the proof.

To prove existence of the solution define

$$\begin{aligned} x_{1,n+1}(t) &= x_{10} + \int_{t_0}^{t} f_1(x_{1,n}(t), x_{2,n}(t), t)\, dt \\ x_{2,n+1}(t) &= x_{20} + \int_{t_0}^{t} f_2(x_{1,n}(t), x_{2,n}(t), t)\, dt \\ &(n = 0, 1, 2, \ldots) \end{aligned}$$



(A)



(B)

Figure 1-16 On-off control system. (A) Block diagram; (B) analog computer diagram.

where $x_{1,0}(t) = x_{10}, \quad x_{2,0}(t) = x_{20}$

and show convergence of the series, obtain $x_{1,n}(t)$ and $x_{2,n}(t)$ and then let n approach infinity. This will result in

$$x_1(t) = x_{10} + \int_{t_0}^{t} f_1(x_1(t), x_2(t), t)\, dt$$

$$x_2(t) = x_{20} + \int_{t_0}^{t} f_2(x_1(t), x_2(t), t)\, dt$$

To prove uniqueness of the solution, assume $\hat{x}_1(t)$ and $\hat{x}_2(t)$ as another set of solution of Eq. (1–31). Then prove that $\hat{x}_1(t) \longrightarrow x_1(t)$, and $\hat{x}_2(t) \longrightarrow x_2(t)$ as $n \longrightarrow \infty$.

The proof can, of course, be extended to the system of n first-order differential equations.

Problem A-1-13 Obtain a state space equation of the on-off control system shown in Fig. 1–16(A). Assume that we are interested in the response of the system subjected to the initial condition only.

Solution The differential equation of the plant is

$$\ddot{x} + a\dot{x} = Ku \tag{1–32}$$

where $u = -\operatorname{sgn} x$

By defining

$$x_1 = x, \qquad x_2 = \dot{x}$$

Eq. (1–32) can be written as

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -K \operatorname{sgn} x_1 - ax_2 \end{aligned} \tag{1–33}$$

Notice that $-K \operatorname{sgn} x_1 - ax_2$ is a piecewise continuous function having discontinuities whenever x_1 changes its sign. Since $-K \operatorname{sgn} x_1 - ax_2$ involves only discontinuities of the magnitude $2K$, Eq. (1–33) possesses a unique solution satisfying the given initial condition. Hence x_1 and x_2 are a set of state variables and Eq. (1–33) is a state space representation of the system.

Figure 1–16(B) shows an analog computer diagram for this system. Observe that the output signals from the integrators can be a set of state variables.

PROBLEMS

Problem B-1-1 A particle of mass m moves under the gravitational field. The gravitational force is defined by the following potential function:

$$V = -m\frac{k^2}{r}$$

where r is the distance from the origin of the gravitational field and k is a constant. By using polar coordinates r, θ, obtain the kinetic energy T of the particle. Then define

$$q_1 = r, \qquad q_2 = \theta$$

$$p_1 = \frac{\partial T}{\partial \dot{q}_1}, \qquad p_2 = \frac{\partial T}{\partial \dot{q}_2}$$

and derive Hamilton's equations of motion of the particle.

Problem B-1-2 Consider the system shown in Fig. 1–17. By defining a suitable set of state variables obtain a state space equation of the system.

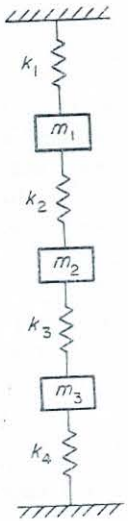


Figure 1-17 Mechanical vibratory system.

Problem B-1-3 Consider the following system:

$$\ddot{x} + f(x)\dot{x} + g(x) = 0$$

Define

$$F(x) = \int_0^x f(\sigma)\, d\sigma$$

and $x_1 = x$, $x_2 = \dot{x} + F(x)$. Show that the given system can be written in the following form:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{bmatrix} = \begin{bmatrix} x_2 - F(x_1) \\ -g(x_1) \end{bmatrix}$$

Problem B-1-4 Derive a state space equation of the system shown in Fig. 1–18. Here, x_1 and x_2 are a set of state variables and u_1 and u_2 are control variables.

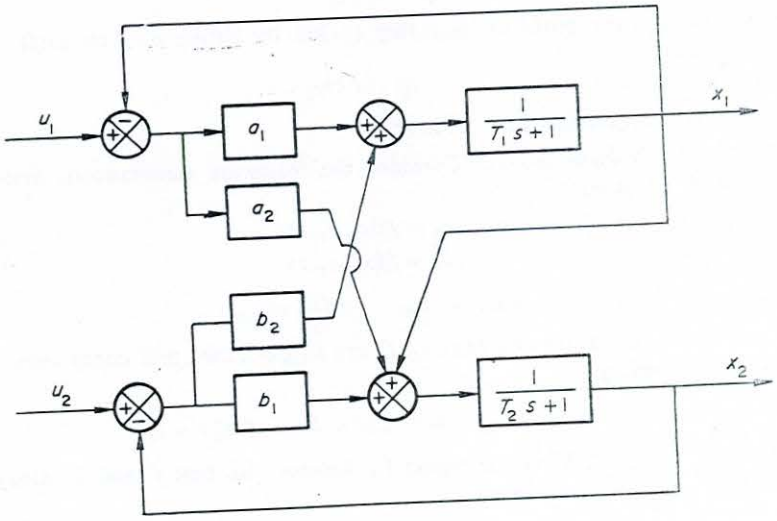


Figure 1-18 Multiple-input multiple-output system.

Problem B-1-5 Figure 1–19 shows a simple hydraulic system. Derive a state space equation of the system. Take

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} x(t) \\ \dot{x}(t) \end{bmatrix}$$

as a state vector. In this system $u(t)$ is the control function. Assume that oil leakage, compressibility including the effects of dissolved air, expansion of pipelines, etc. are negligible.

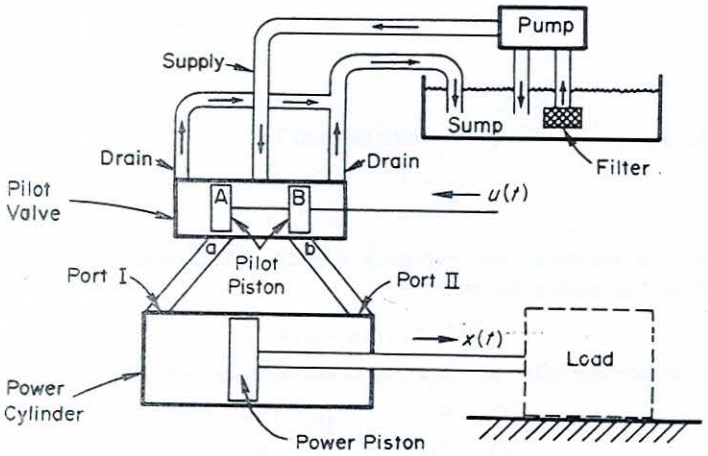


Figure 1-19 Hydraulic system.

Note that if we define

S = power piston area in.^{2}
$Q = S\dot{x}$ = rate of flow of fluid to the power cylinder, in.^{3}/sec
ΔP = pressure difference across the power piston, psi
u = displacement of pilot valve, in. (control function)

then Q can be given by

$$Q = K_1 u - K_2 \Delta P$$

where K_1 and K_2 are constants.

REFERENCES

1–1 Enns, M., J. R. Greenwood III, J. E. Matheson, and F. T. Thompson, "Practical Aspects of State Space Methods: Part 1—System Formulation and Reduction," *JACC* Preprint (1964) 494–500.

1–2 Fuller, A. T., "Phase Space in the Theory of Optimum Control," *J. Elect. Cont.*, 8, No. 5 (1960), 381–400.

1–3 Dertouzos, M. L., and J. K. Roberge, "High Capacity Reaction-Wheel Attitude Control," *JACC* Preprint (1963) 428–36.

2

Review of Matrices and Vectors

2-1 INTRODUCTION

The principal objective of this chapter is to review elementary vector matrix analysis.[1] Section 2–1 presents mainly definitions of matrices and associated terminologies. Section 2–2 gives some of the rules of matrix operations. Section 2–3 discusses briefly vectors and vector spaces. Finally Section 2–4 presents definitions of terminologies commonly appearing in matrix analysis. Discussions of matrix polynominals, Cayley-Hamilton theorem, the minimal polynomial, the exponential function of a matrix, etc. are postponed to Chapter 5.

We now give definitions of matrices and associated terminologies:

Number fields A set F of numbers containing at least two numbers is a *field* if a, b are any two members of F, then $a + b$, $a - b$, ab, a/b (where $b \neq 0$) are also members of F. For example, the set of all real numbers constitutes a field, and so does that of all complex numbers. The set of integers, however, is not a field, since a/b (where $b \neq 0$) may not be a member of F. Similarly, neither the set of positive numbers nor the set of negative numbers is a field.

[1] Readers interested in details of vector matrix analysis refer to specialized works, for example (*2–1*, *2–2*, *2–3*, *2–4* among many others). The figures in parentheses refer to books listed at the end of this chapter.

In this book the fields that concern us are the real field (the set of all real numbers) and the complex field (the set of all complex numbers.)

Matrix A matrix $\mathbf{A}$ over a field F is a rectangular array of numbers in F. If $\mathbf{A}$ is over the field of real numbers, $\mathbf{A}$ is called a *real* matrix. If $\mathbf{A}$ is over the field of complex numbers, $\mathbf{A}$ is called a *complex* matrix.

In this book we may not mention each time the field to which the elements of the matrix belong, but we shall assume that such a field has been given.

In a matrix the number of columns, in general, is not necessarily the same as that of rows. Consider the following matrix:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} \tag{2-1}$$

The number standing in the ith row and jth column is denoted by a_{ij}. The numbers a_{ij} are called the *elements* of $\mathbf{A}$. This matrix has n rows and m columns and is called an $n \times m$ matrix. The first suffix represents the row number and the second suffix the column number. The matrix $\mathbf{A}$ given by Eq. (2–1) is sometimes written as (a_{ij}), where we indicate only the (i, j)th element. Two matrices are called *equal* if and only if their corresponding elements are equal.

Matrices are closely related to linear transformations. Consider a system of n linear equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m &= z_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2m}x_m &= z_2 \\ \cdots \quad \cdots \quad \cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nm}x_m &= z_n \end{aligned}$$

This set of n equations assigns a unique set of n variables $z_1, z_2, \ldots, z_n$ to every set of m variables $x_1, x_2, \ldots, x_m$. Such an assignment is called a *linear transformation* of the set of variables $x_1, x_2, \ldots, x_m$ into the set of variables $z_1, z_2, \ldots, z_n$. This linear transformation is characterized by the array of nm coefficients a_{ij} as given by Eq. (2–1). Hence $\mathbf{A} = (a_{ij})$ is called the *matrix of the linear transformation.*

Square matrix A *square matrix* is a matrix in which the number of rows is equal to that of columns. Such a square matrix is called a matrix of order n, where n is the number of rows or columns. It is commonly called an $n \times n$ matrix.

Trace The trace of an $n \times n$ matrix $\mathbf{A}$ is defined by

$$\text{trace of } \mathbf{A} = \text{tr } \mathbf{A} = \sum_{i=1}^{n} a_{ii}$$

Diagonal matrix The $n \times n$ matrix having the elements $a_{11}, a_{22}, \ldots, a_{nn}$ in its main diagonal and zeros in all other positions is called a *diagonal* matrix and is written as

$$\mathbf{A} = \begin{bmatrix} a_{11} & & & 0 \\ & a_{22} & & \\ & & \ddots & \\ 0 & & & a_{nn} \end{bmatrix} = (a_{ij}\,\delta_{ij})$$

where the δ_{ij} are the *Kronecker delta* defined by

$$\begin{aligned} \delta_{ij} &= 1 \qquad (i = j) \\ &= 0 \qquad (i \neq j) \end{aligned}$$

(Note that all the terms not explicitly written in the matrix are zero.) The diagonal matrix is also denoted by

$$\text{diag } (a_{11}, a_{22}, \ldots, a_{nn})$$

These notations are often extended to cover submatrices. For example, if

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} a_{33} & a_{34} \\ a_{43} & a_{44} \end{bmatrix}$$

then the matrix

$$\left[\begin{array}{cc|cc} a_{11} & a_{12} & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 \\ \hline 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & a_{43} & a_{44} \end{array}\right]$$

may be denoted by

$$\begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix} \quad \text{or} \quad \text{diag } (\mathbf{A}, \mathbf{B})$$

A matrix whose elements are divided into submatrices by vertical and horizontal dotted lines is called a *partitioned* matrix. We shall give a detailed discussion of partitioned matrices later.

Identity matrix or unity matrix The *identity* matrix or *unity* matrix $\mathbf{I}$ is a matrix whose elements on the main diagonal are equal to unity and all other elements are equal to zero,

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \text{diag}\,(1, 1, \ldots, 1)$$

The identity matrix $\mathbf{I}$ is sometimes written as $\mathbf{I} = (\delta_{ij})$ where the δ_{ij} are the Kronecker delta. If it is necessary to indicate the number of rows (or columns) of the identity matrix we shall put the suffix corresponding to the number of rows (or columns). For example, an $n \times n$ identity matrix will be denoted by $\mathbf{I}_n$.

The identity matrix $\mathbf{I}$ is uniquely determined by the relation $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$ for all square matrices $\mathbf{A}$. Therefore, if $\mathbf{AB} = \mathbf{BA}$ for all $\mathbf{B}$, then $\mathbf{A}$ is a scalar multiple of $\mathbf{I}$. (For multiplication of a matrix by a matrix, see Section 2–2.)

Scalar matrix A diagonal matrix $\mathbf{A}$ whose elements are equal to one another is called a *scalar* matrix.

Triangular matrix A matrix whose elements below (or above) the main diagonal are all zero is called a *triangular* matrix. Examples are

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & a_{44} \end{bmatrix}, \quad \begin{bmatrix} a_{11} & 0 & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 \\ a_{31} & a_{32} & a_{33} & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

Null matrix The null matrix $\mathbf{0}$ is determined uniquely by the relation $\mathbf{A0} = \mathbf{0A} = \mathbf{0}$ for all $\mathbf{A}$.

Transpose If the rows and columns of an $n \times m$ matrix $\mathbf{A}$ are interchanged, the resulting $m \times n$ matrix is called the *transpose* of $\mathbf{A}$. The transpose of $\mathbf{A}$ is denoted by $\mathbf{A}'$. Namely, if $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

then $\mathbf{A}'$ is given by

$$\mathbf{A}' = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{bmatrix}$$

Note that $(\mathbf{A}')' = \mathbf{A}$. We can easily verify that if $\mathbf{A} + \mathbf{B}$ and $\mathbf{AB}$ can be defined (see Section 2–2), then

$$(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}', \qquad (\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$$

Conjugate matrix The *conjugate* of a matrix $\mathbf{A}$ is the matrix whose each element is the conjugate complex of the corresponding element of $\mathbf{A}$. The conjugate of $\mathbf{A}$ is denoted by $\bar{\mathbf{A}} = (\bar{a}_{ij})$, where $\bar{a}_{ij}$ is the complex conjugate of a_{ij}. For example, if $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ -1+j & -3-3j & -1+4j \\ -1+j & -1 & -2+3j \end{bmatrix} \tag{2–2}$$

then

$$\bar{\mathbf{A}} = \begin{bmatrix} 0 & 1 & 0 \\ -1-j & -3+3j & -1-4j \\ -1-j & -1 & -2-3j \end{bmatrix}$$

Conjugate transpose The conjugate transpose is the conjugate of the transpose of a matrix. Given a matrix $\mathbf{A} = (a_{ij})$, the conjugate transpose is denoted by $\bar{\mathbf{A}}'$ or $\mathbf{A}^*$, namely,

$$\bar{\mathbf{A}}' = \mathbf{A}^* = (\bar{a}_{ji})$$

For example, if $\mathbf{A}$ is given by Eq. (2–2), then

$$\bar{\mathbf{A}}' = \mathbf{A}^* = \begin{bmatrix} 0 & -1-j & -1-j \\ 1 & -3+3j & -1 \\ 0 & -1-4j & -2-3j \end{bmatrix}$$

Clearly, the conjugate of $\mathbf{A}'$ is the same as the transpose of $\bar{\mathbf{A}}$. Note that $(\mathbf{A}^*)^* = \mathbf{A}$. We can easily verify that if $\mathbf{A} + \mathbf{B}$ and $\mathbf{AB}$ can be defined, then

$$(\mathbf{A} + \mathbf{B})^* = \mathbf{A}^* + \mathbf{B}^*, \qquad (\mathbf{AB})^* = \mathbf{B}^*\mathbf{A}^*$$

Note also that if c is a complex number, then

$$(c\mathbf{A})^* = \bar{c}\mathbf{A}^*$$

If $\mathbf{A}$ is a real matrix (a matrix whose elements are real), the conjugate transpose $\mathbf{A}^*$ is the same as the transpose $\mathbf{A}'$.

m-rowed minor An m-rowed determinant obtainable by deleting $n - m$ rows and $n - m$ columns from an n-rowed determinant is called an *m-rowed* minor.

Minor M_{ij} We shall employ the notation M_{ij} for an $(n - 1)$-

rowed determinant obtainable by deleting the ith row and jth column from an n-rowed determinant.

Principal minor A minor of a determinant is called a *principal minor*, if its diagonal elements are also diagonal elements of the determinant.

Cofactor Δ_{ij} The cofactor Δ_{ij} is defined by $\Delta_{ij} = (-)^{i+j}M_{ij}$. The cofactor Δ_{ij} is the coefficient of the term a_{ij} in the expansion of $|\mathbf{A}|$, the determinant of $\mathbf{A}$, by minors, or

$$a_{i1}\Delta_{i1} + a_{i2}\Delta_{i2} + \cdots + a_{in}\Delta_{in} = |\mathbf{A}|$$

In this book we denote the determinant of $\mathbf{A}$ by $|\mathbf{A}|$ or det $\mathbf{A}$. If $a_{i1}, a_{i2}, \ldots, a_{in}$ are replaced by $a_{j1}, a_{j2}, \ldots, a_{jn}$, then

$$a_{j1}\Delta_{i1} + a_{j2}\Delta_{i2} + \cdots + a_{jn}\Delta_{in} = 0 \qquad (i \neq j)$$

because $\mathbf{A}$ can be seen to possess two identical rows. Hence we obtain

$$\sum_{k=1}^{n} a_{jk}\Delta_{ik} = \delta_{ji}|\mathbf{A}| \tag{2–3}$$

Similarly,

$$\sum_{k=1}^{n} a_{ki}\Delta_{kj} = \delta_{ij}|\mathbf{A}| \tag{2–4}$$

Adjoint of square matrix The *adjoint* of a square matrix $\mathbf{A}$ is the transpose of the matrix obtained from $\mathbf{A}$ by replacing each element of $\mathbf{A}$ by its cofactor. Namely, the adjoint of a square matrix $\mathbf{A}$ is a matrix $\mathbf{B} = (b_{ij})$, where $b_{ij} = \Delta_{ji}$. The adjoint of a square matrix $\mathbf{A}$ is written as adj $\mathbf{A}$. The element in the jth row and ith column of the product $\mathbf{A}$(adj $\mathbf{A}$) is

$$\sum_{k=1}^{n} a_{jk}b_{ki} = \sum_{k=1}^{n} a_{jk}\Delta_{ik} = \delta_{ji}|\mathbf{A}|$$

Hence $\mathbf{A}$(adj $\mathbf{A}$) is a diagonal matrix with diagonal elements equal to $|\mathbf{A}|$. Thus $\mathbf{A}$(adj $\mathbf{A}$) $= |\mathbf{A}|\mathbf{I}$. Similarly, the element in the jth row and the ith column of the product (adj $\mathbf{A}$) $\mathbf{A}$ is

$$\sum_{k=1}^{n} b_{jk}a_{ki} = \sum_{k=1}^{n} \Delta_{kj}a_{ki} = \delta_{ij}|\mathbf{A}|$$

Hence we have the relationship

$$\mathbf{A}(\text{adj } \mathbf{A}) = (\text{adj } \mathbf{A})\mathbf{A} = |\mathbf{A}|\mathbf{I} \tag{2–5}$$

For example, given a matrix $\mathbf{A}$

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 0 \\ 3 & -1 & -2 \\ 1 & 0 & -3 \end{bmatrix}$$

adj $\mathbf{A}$ is obtained as

$$\text{adj } \mathbf{A} = \begin{bmatrix} \begin{vmatrix} -1 & -2 \\ 0 & -3 \end{vmatrix} & -\begin{vmatrix} 2 & 0 \\ 0 & -3 \end{vmatrix} & \begin{vmatrix} 2 & 0 \\ -1 & -2 \end{vmatrix} \\ -\begin{vmatrix} 3 & -2 \\ 1 & -3 \end{vmatrix} & \begin{vmatrix} 1 & 0 \\ 1 & -3 \end{vmatrix} & -\begin{vmatrix} 1 & 0 \\ 3 & -2 \end{vmatrix} \\ \begin{vmatrix} 3 & -1 \\ 1 & 0 \end{vmatrix} & -\begin{vmatrix} 1 & 2 \\ 1 & 0 \end{vmatrix} & \begin{vmatrix} 1 & 2 \\ 3 & -1 \end{vmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 6 & -4 \\ 7 & -3 & 2 \\ 1 & 2 & -7 \end{bmatrix}$$

and

$$\mathbf{A}(\text{adj } \mathbf{A}) = \begin{bmatrix} 1 & 2 & 0 \\ 3 & -1 & -2 \\ 1 & 0 & -3 \end{bmatrix}\begin{bmatrix} 3 & 6 & -4 \\ 7 & -3 & 2 \\ 1 & 2 & -7 \end{bmatrix} = 17\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Nonsingular matrix and singular matrix A square matrix $\mathbf{A}$ is called *nonsingular* if a matrix $\mathbf{B}$ exists such that $\mathbf{BA} = \mathbf{AB} = \mathbf{I}$. If such a matrix $\mathbf{B}$ exists, then it is denoted by $\mathbf{A}^{-1}$. $\mathbf{A}^{-1}$ is called the *inverse* of $\mathbf{A}$. The inverse matrix $\mathbf{A}^{-1}$ exists if $|\mathbf{A}|$ is non-zero. If $\mathbf{A}^{-1}$ does not exist, $\mathbf{A}$ is called *singular*.

Inverse matrix The inverse matrix $\mathbf{A}^{-1}$, which exists for a nonsingular matrix $\mathbf{A}$, has the property that $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix. Thus, $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$. If $\mathbf{AB} = \mathbf{C}$ then, $\mathbf{A}^{-1}\mathbf{AB} = \mathbf{IB} = \mathbf{B} = \mathbf{A}^{-1}\mathbf{C}$. If $\mathbf{A}$ and $\mathbf{B}$ are nonsingular matrices, then the product $\mathbf{AB}$ is a nonsingular matrix. Moreover, $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$. This may be seen as follows:

$$(\mathbf{B}^{-1}\mathbf{A}^{-1})\mathbf{AB} = \mathbf{B}^{-1}(\mathbf{A}^{-1}\mathbf{A})\mathbf{B} = \mathbf{B}^{-1}\mathbf{IB} = \mathbf{B}^{-1}\mathbf{B} = \mathbf{I}$$

Similarly, we obtain $(\mathbf{AB})(\mathbf{B}^{-1}\mathbf{A}^{-1}) = \mathbf{I}$. Hence we verify that

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

Notice that for a nonsingular matrix $\mathbf{A}$,

$$\mathbf{A}^{-k} = (\mathbf{AA}\cdots\mathbf{A})^{-1} = \mathbf{A}^{-1}\mathbf{A}^{-1}\cdots\mathbf{A}^{-1} = (\mathbf{A}^{-1})^{k}$$

Also

$$(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$$

This can be proved by taking the transpose of

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

Namely,

$$(\mathbf{A}^{-1})'\mathbf{A}' = \mathbf{A}'(\mathbf{A}^{-1})' = \mathbf{I}' = \mathbf{I}$$

it follows that $(\mathbf{A}^{-1})'$ is the inverse of $\mathbf{A}'$, or

$$(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$$

Similarly, we have the following relationship:

$$(\mathbf{A}^*)^{-1} = (\mathbf{A}^{-1})^*$$

From Eq. (2–5) and the definition of the inverse of a matrix, we have

$$\mathbf{A}^{-1} = \frac{\text{adj } \mathbf{A}}{|\mathbf{A}|}$$

Hence the *inverse* of a nonsingular matrix is the transpose of the matrix of its cofactors divided by the determinant of the matrix. If $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

then

$$\mathbf{A}^{-1} = \frac{\text{adj } \mathbf{A}}{|\mathbf{A}|} = \begin{bmatrix} \frac{\Delta_{11}}{|\mathbf{A}|} & \frac{\Delta_{21}}{|\mathbf{A}|} & \cdots & \frac{\Delta_{n1}}{|\mathbf{A}|} \\ \frac{\Delta_{12}}{|\mathbf{A}|} & \frac{\Delta_{22}}{|\mathbf{A}|} & \cdots & \frac{\Delta_{n2}}{|\mathbf{A}|} \\ \vdots & \vdots & & \vdots \\ \frac{\Delta_{1n}}{|\mathbf{A}|} & \frac{\Delta_{2n}}{|\mathbf{A}|} & \cdots & \frac{\Delta_{nn}}{|\mathbf{A}|} \end{bmatrix}$$

where Δ_{ij} is the cofactor of a_{ij} of $\mathbf{A}$. Hence the terms in the ith column of $\mathbf{A}^{-1}$ are $1/|\mathbf{A}|$ times the cofactors of the ith row of the original matrix $\mathbf{A}$. For example, if $\mathbf{A}$ is given as

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 0 \\ 3 & -1 & -2 \\ 1 & 0 & -3 \end{bmatrix}$$

then adj $\mathbf{A}$ and $|\mathbf{A}|$ are obtained respectively as

$$\text{adj } \mathbf{A} = \begin{bmatrix} 3 & 6 & -4 \\ 7 & -3 & 2 \\ 1 & 2 & -7 \end{bmatrix}, \qquad |\mathbf{A}| = 17$$

Hence the inverse of $\mathbf{A}$ is given by

$$\mathbf{A}^{-1} = \frac{\text{adj } \mathbf{A}}{|\mathbf{A}|} = \begin{bmatrix} \frac{3}{17} & \frac{6}{17} & -\frac{4}{17} \\ \frac{7}{17} & -\frac{3}{17} & \frac{2}{17} \\ \frac{1}{17} & \frac{2}{17} & -\frac{7}{17} \end{bmatrix}$$

Formulas for finding inverse matrices are given below for a 2×2 matrix and 3×3 matrix. For a 2×2 matrix $\mathbf{A}$ where

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \qquad ad - bc \neq 0$$

the inverse matrix is given by

$$\mathbf{A}^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

For a 3×3 matrix $\mathbf{A}$ where

$$\mathbf{A} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}, \qquad |\mathbf{A}| \neq 0$$

the inverse matrix is given by

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|}\begin{bmatrix} \begin{vmatrix} e & f \\ h & i \end{vmatrix} & -\begin{vmatrix} b & c \\ h & i \end{vmatrix} & \begin{vmatrix} b & c \\ e & f \end{vmatrix} \\ -\begin{vmatrix} d & f \\ g & i \end{vmatrix} & \begin{vmatrix} a & c \\ g & i \end{vmatrix} & -\begin{vmatrix} a & c \\ d & f \end{vmatrix} \\ \begin{vmatrix} d & e \\ g & h \end{vmatrix} & -\begin{vmatrix} a & b \\ g & h \end{vmatrix} & \begin{vmatrix} a & b \\ d & e \end{vmatrix} \end{bmatrix}$$

The inverse of $\mathbf{A}$ can also be obtained by a different method which is given now. (Another method is given in Section 5–2.)

Consider n simultaneous equations in $x_1, x_2, \ldots, x_n$.

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \cdots \qquad \cdots \qquad &\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned}$$

or in the form of the vector matrix equation

$$\mathbf{A}\mathbf{x} = \mathbf{B}$$

Solving for $\mathbf{x}$, we obtain

$$\begin{aligned} x_1 &= c_{11}b_1 + c_{12}b_2 + \cdots + c_{1n}b_n \\ x_2 &= c_{21}b_1 + c_{22}b_2 + \cdots + c_{2n}b_n \\ &\cdots \quad \cdots \quad \cdots \\ x_n &= c_{n1}b_1 + c_{n2}b_2 + \cdots + c_{nn}b_n \end{aligned}$$

or

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

which is equivalent to

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{B}$$

Hence

$$\mathbf{A}^{-1} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix}$$

This method is practical if it is possible to obtain $x_1, x_2, \ldots, x_n$ in terms of $b_1, b_2, \ldots, b_n$ without much computational efforts.

Symmetric matrix and skew symmetric matrix A *symmetric* matrix is a matrix which is equal to its transpose. Namely, for a symmetric matrix $\mathbf{A}$

$$\mathbf{A}' = \mathbf{A} \qquad \text{or} \qquad a_{ji} = a_{ij}$$

If a matrix $\mathbf{A}$ is equal to the negative of its transpose, namely,

$$\mathbf{A}' = -\mathbf{A} \qquad \text{or} \qquad a_{ji} = -a_{ij}$$

then it is called a *skew symmetric matrix*.

If $\mathbf{A}$ is any square matrix, then $\mathbf{A} + \mathbf{A}'$ is a symmetric matrix and $\mathbf{A} - \mathbf{A}'$ is a skew symmetric matrix. For example, if $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

then

$$\mathbf{A} + \mathbf{A}' = \begin{bmatrix} 2 & 6 & 10 \\ 6 & 10 & 14 \\ 10 & 14 & 18 \end{bmatrix} = \text{symmetric}$$

and

$$\mathbf{A} - \mathbf{A}' = \begin{bmatrix} 0 & -2 & -4 \\ 2 & 0 & -2 \\ 4 & 2 & 0 \end{bmatrix} = \text{skew symmetric}$$

Notice that if $\mathbf{A}$ is a rectangular matrix, then $\mathbf{A}'\mathbf{A}$ is a symmetric matrix. Notice also that the inverse of a symmetric matrix is symmetric. To establish this fact take the transpose of $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. We have $(\mathbf{A}^{-1})'\mathbf{A}' = \mathbf{I}' = \mathbf{I}$. Noting that $\mathbf{A} = \mathbf{A}'$, we have $(\mathbf{A}^{-1})'\mathbf{A}' = (\mathbf{A}^{-1})'\mathbf{A} = \mathbf{I} = \mathbf{A}^{-1}\mathbf{A}$. Hence, $\mathbf{A}^{-1} = (\mathbf{A}^{-1})'$. Thus, the inverse of a symmetric matrix is symmetric.

Orthogonal matrix A matrix $\mathbf{A}$ is an *orthogonal* matrix if it is real and satisfies the relationship $\mathbf{A}'\mathbf{A} = \mathbf{A}\mathbf{A}' = \mathbf{I}$. (This implies that $|\mathbf{A}| = \pm 1$ and $\mathbf{A}$ is nonsingular.) In the orthogonal matrix the inverse is exactly equal to its transpose.

$$\mathbf{A}^{-1} = \mathbf{A}'$$

Examples of orthogonal matrices are

$$\mathbf{A} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0.6 & 0.8 & 0 \\ -0.8 & 0.6 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

To demonstrate, let us compute $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$

$$\mathbf{A}^{-1} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} = \mathbf{A}', \qquad \mathbf{B}^{-1} = \begin{bmatrix} 0.6 & -0.8 & 0 \\ 0.8 & 0.6 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{B}'$$

If $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ orthogonal matrices, then so are $\mathbf{A}^{-1}$, $\mathbf{A}'$, and $\mathbf{AB}$. This may be seen as follows: Since $\mathbf{A}$ is orthogonal, $\mathbf{A}\mathbf{A}' = \mathbf{I}$ and $(\mathbf{A}')'\mathbf{A}' = \mathbf{I}$. Hence $\mathbf{A}'$ is orthogonal. Since $\mathbf{A}^{-1} = \mathbf{A}'$, $\mathbf{A}^{-1}$ is also orthogonal. Since $\mathbf{B}' = \mathbf{B}^{-1}$, $\mathbf{A}' = \mathbf{A}^{-1}$, $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$ and $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$, we have $(\mathbf{AB})' = (\mathbf{AB})^{-1}$, so $\mathbf{AB}$ is orthogonal.

Hermitian matrix and skew Hermitian matrix A matrix whose elements are complex quantities is called a *complex* matrix. If a complex matrix $\mathbf{A}$ satisfies the relationship

$$\mathbf{A}^* = \mathbf{A} \qquad \text{or} \qquad a_{ij} = \bar{a}_{ji}$$

where $\bar{a}_{ji}$ is the complex conjugate of a_{ji}, then $\mathbf{A}$ is called a *Hermitian* matrix. An example is

$$\mathbf{A} = \begin{bmatrix} 1 & 4+j3 \\ 4-j3 & 2 \end{bmatrix}$$

If a Hermitian matrix $\mathbf{A}$ is written as $\mathbf{A} = \mathbf{B} + j\mathbf{C}$, where $\mathbf{B}$ and $\mathbf{C}$ are real matrices, then

$$\mathbf{B} = \mathbf{B}', \qquad \mathbf{C} = -\mathbf{C}'$$

In the foregoing example

$$\mathbf{A} = \mathbf{B} + j\mathbf{C} = \begin{bmatrix} 1 & 4 \\ 4 & 2 \end{bmatrix} + j\begin{bmatrix} 0 & 3 \\ -3 & 0 \end{bmatrix}$$

Notice that the inverse of a Hermitian matrix $\mathbf{A}$ is Hermitian, or $\mathbf{A}^{-1} = (\mathbf{A}^{-1})^*$. Notice also that every square matrix can be expressed uniquely as $\mathbf{A} = \mathbf{G} + j\mathbf{H}$, where $\mathbf{G}$ and $\mathbf{H}$ are Hermitian and are given by

$$\mathbf{G} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^*), \qquad \mathbf{H} = \frac{1}{2j}(\mathbf{A} - \mathbf{A}^*)$$

The fact that $\mathbf{G}$ and $\mathbf{H}$ are Hermitian is seen as follows:

$$\mathbf{G}^* = \frac{1}{2}(\mathbf{A}^* + \mathbf{A}) = \mathbf{G}$$

$$\mathbf{H}^* = \frac{-1}{2j}(\mathbf{A}^* - \mathbf{A}) = \mathbf{H}$$

We can easily verify that for $n \times n$ Hermitian matrices $\mathbf{A}$ and $\mathbf{B}$, the matrices $\mathbf{A} + \mathbf{B}$, $\mathbf{A} - \mathbf{B}$, $\mathbf{AB} + \mathbf{BA}$ are also Hermitian. The product $\mathbf{AB}$ is Hermitian, however, if and only if $\mathbf{A}$ and $\mathbf{B}$ commute, since $\mathbf{AB} = \mathbf{A}^*\mathbf{B}^* = (\mathbf{BA})^*$. The determinant of a Hermitian matrix is always real, since

$$|\mathbf{A}| = |\mathbf{A}^*| = |\bar{\mathbf{A}}'| = |\bar{\mathbf{A}}|$$

If a matrix $\mathbf{A}$ satisfies the relationship $\mathbf{A}^* = -\mathbf{A}$, then $\mathbf{A}$ is called a *skew Hermitian* matrix. An example is

$$\mathbf{A} = \begin{bmatrix} j5 & -2+j3 \\ 2+j3 & j \end{bmatrix}$$

If a skew Hermitian matrix $\mathbf{A}$ is written as $\mathbf{A} = \mathbf{B} + j\mathbf{C}$, where $\mathbf{B}$ and $\mathbf{C}$ are real matrices, then

$$\mathbf{B} = -\mathbf{B}', \qquad \mathbf{C} = \mathbf{C}'$$

In the present example,

$$\mathbf{A} = \mathbf{B} + j\mathbf{C} = \begin{bmatrix} 0 & -2 \\ 2 & 0 \end{bmatrix} + j\begin{bmatrix} 5 & 3 \\ 3 & 1 \end{bmatrix}$$

Unitary matrix A *unitary* matrix is a complex matrix in which the inverse is equal to the conjugate of its transpose. Namely,

$$\mathbf{A}^{-1} = \mathbf{A}^* \qquad \text{or} \qquad \mathbf{AA}^* = \mathbf{A}^*\mathbf{A} = \mathbf{I}$$

An example of a unitary matrix is

$$\mathbf{A} = \begin{bmatrix} \frac{1}{\sqrt{15}}(2+j) & \frac{1}{\sqrt{15}}(3+j) \\ \frac{1}{\sqrt{15}}(-3+j) & \frac{1}{\sqrt{15}}(2-j) \end{bmatrix}$$

To demonstrate, let us compute $\mathbf{A}^{-1}$ and $\mathbf{A}^*$. Since for a unitary matrix $\mathbf{A}$ the determinant of $\mathbf{A}$ is equal to one or $|\mathbf{A}| = 1$, we obtain

$$\mathbf{A}^{-1} = \begin{bmatrix} \frac{1}{\sqrt{15}}(2-j) & -\frac{1}{\sqrt{15}}(3+j) \\ \frac{1}{\sqrt{15}}(3-j) & \frac{1}{\sqrt{15}}(2+j) \end{bmatrix}$$

The conjugate transpose $\mathbf{A}^*$ is given by

$$\mathbf{A}^* = \begin{bmatrix} \frac{1}{\sqrt{15}}(2-j) & -\frac{1}{\sqrt{15}}(3+j) \\ \frac{1}{\sqrt{15}}(3-j) & \frac{1}{\sqrt{15}}(2+j) \end{bmatrix}$$

Hence we verify that $\mathbf{A}^{-1} = \mathbf{A}^*$.

Orthogonal matrices satisfy the relationship, $\mathbf{AA}^* = \mathbf{A}^*\mathbf{A} = \mathbf{I}$; hence they are unitary.

Notice that if $\mathbf{A}$ is unitary, then the inverse $\mathbf{A}^{-1}$ is also unitary. To see this note that since $\mathbf{AA}^* = \mathbf{A}^*\mathbf{A} = \mathbf{I}$, we obtain

$$(\mathbf{A}^*)^{-1}\mathbf{A}^{-1} = (\mathbf{A}^{-1})^*(\mathbf{A}^{-1}) = \mathbf{I}$$

$$\mathbf{A}^{-1}(\mathbf{A}^*)^{-1} = (\mathbf{A}^{-1})(\mathbf{A}^{-1})^* = \mathbf{I}$$

If $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ are unitary, then $\mathbf{AB}$ is also unitary. To prove this, notice that since $\mathbf{AA}^* = \mathbf{A}^*\mathbf{A} = \mathbf{I}$ and $\mathbf{BB}^* = \mathbf{B}^*\mathbf{B} = \mathbf{I}$, we have

$$(\mathbf{AB})(\mathbf{AB})^* = \mathbf{ABB}^*\mathbf{A}^* = \mathbf{AA}^* = \mathbf{I}$$

$$(\mathbf{AB})^*(\mathbf{AB}) = \mathbf{B}^*\mathbf{A}^*\mathbf{AB} = \mathbf{B}^*\mathbf{B} = \mathbf{I}$$

Hence it is shown that $\mathbf{AB}$ is unitary.

Normal matrix A matrix which commutes with its conjugate transpose is called a *normal* matrix. Namely, for a normal matrix $\mathbf{A}$

$$\mathbf{AA}^* = \mathbf{A}^*\mathbf{A} \qquad \text{(if } \mathbf{A} \text{ is a complex matrix)}$$
$$\mathbf{AA}' = \mathbf{A}'\mathbf{A} \qquad \text{(if } \mathbf{A} \text{ is a real matrix)}$$

Notice that if $\mathbf{A}$ is normal and $\mathbf{U}$ is unitary, then $\mathbf{U}^{-1}\mathbf{AU}$ is also normal, since

$$(\mathbf{U}^{-1}\mathbf{AU})(\mathbf{U}^{-1}\mathbf{AU})^* = \mathbf{U}^{-1}\mathbf{AUU}^*\mathbf{A}^*(\mathbf{U}^{-1})^* = \mathbf{U}^{-1}\mathbf{AA}^*(\mathbf{U}^{-1})^*$$
$$= \mathbf{U}^*\mathbf{A}^*\mathbf{AU} = \mathbf{U}^*\mathbf{A}^*(\mathbf{U}^{-1})^*\mathbf{U}^{-1}\mathbf{AU} = (\mathbf{U}^{-1}\mathbf{AU})^*(\mathbf{U}^{-1}\mathbf{AU})$$

A matrix is clearly normal if it is symmetric, or Hermitian, or skew symmetric, or skew Hermitian, or unitary, or orthogonal.

Summary In summarizing definitions of various matrices, we give the following list for a handy reference:

$\mathbf{A}' = \mathbf{A}$	$\mathbf{A}$ is symmetric
$\mathbf{A}' = -\mathbf{A}$	$\mathbf{A}$ is skew symmetric
$\mathbf{AA}' = \mathbf{A}'\mathbf{A} = \mathbf{I}$	$\mathbf{A}$ is orthogonal
$\mathbf{A}^* = \mathbf{A}$	$\mathbf{A}$ is Hermitian
$\mathbf{A}^* = -\mathbf{A}$	$\mathbf{A}$ is skew Hermitian
$\mathbf{AA}^* = \mathbf{A}^*\mathbf{A} = \mathbf{I}$	$\mathbf{A}$ is unitary
$\mathbf{AA}^* = \mathbf{A}^*\mathbf{A}$ or $\mathbf{AA}' = \mathbf{A}'\mathbf{A}$	$\mathbf{A}$ is normal

2-2 RULES OF MATRIX OPERATIONS

This section first reviews some rules of algebraic operations of matrices and then gives definitions of the derivative and integral of matrices. Finally, rules of differentiation of matrices are presented. It will be seen that matrix algebra differs from usual number algebra in that matrix multiplication is not commutative and that cancellation of matrices is not valid.

Sum of matrices The sum of two $n \times m$ matrices $\mathbf{A}$ and $\mathbf{B}$ is a matrix in which each element is the sum of the corresponding elements of two matrices $\mathbf{A}$ and $\mathbf{B}$. Namely,

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1m} + b_{1m} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2m} + b_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} + b_{n1} & a_{n2} + b_{n2} & \cdots & a_{nm} + b_{nm} \end{bmatrix}$$

Note that the sum of matrices $\mathbf{A}$ and $\mathbf{B}$ is not defined unless the number of the rows of $\mathbf{A}$ is the same as that of $\mathbf{B}$ and the number of the columns of $\mathbf{A}$ is the same as that of $\mathbf{B}$.

Difference of matrices If $\mathbf{A}$ and $\mathbf{B}$ are $n \times m$ matrices, the equation $\mathbf{B} + \mathbf{X} = \mathbf{A}$ has a unique solution for $\mathbf{X}$. This solution, $\mathbf{X} = \mathbf{A} - \mathbf{B}$, is called the *difference* of $\mathbf{A}$ and $\mathbf{B}$. The difference $\mathbf{A} - \mathbf{B}$ is given by

$$\mathbf{A} - \mathbf{B} = \begin{bmatrix} a_{11} - b_{11} & a_{12} - b_{12} & \cdots & a_{1m} - b_{1m} \\ a_{21} - b_{21} & a_{22} - b_{22} & \cdots & a_{2m} - b_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} - b_{n1} & a_{n2} - b_{n2} & \cdots & a_{nm} - b_{nm} \end{bmatrix}$$

The difference of matrices $\mathbf{A}$ and $\mathbf{B}$ is not defined unless the number of the rows of $\mathbf{A}$ is equal to that of $\mathbf{B}$ and the number of the columns of $\mathbf{A}$ is equal to that of $\mathbf{B}$.

Multiplication of matrix by scalar The product of a matrix and a scalar is a matrix in which each element is multiplied by the scalar. Namely,

$$k\mathbf{A} = \begin{bmatrix} ka_{11} & ka_{12} & \cdots & ka_{1m} \\ ka_{21} & ka_{22} & \cdots & ka_{2m} \\ \vdots & \vdots & & \vdots \\ ka_{n1} & ka_{n2} & \cdots & ka_{nm} \end{bmatrix}$$

Multiplication of matrix by matrix Multiplication of a matrix by a matrix is possible between matrices in which the number of the columns in the first matrix is equal to the number of the rows in the second. Otherwise, multiplication is not defined.

$$\mathbf{AB} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1r} \\ b_{21} & b_{22} & \cdots & b_{2r} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mr} \end{bmatrix}$$
$$= \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1r} \\ c_{21} & c_{22} & \cdots & c_{2r} \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nr} \end{bmatrix}$$

where $$c_{ik} = \sum_{j=1}^{m} a_{ij}b_{jk}$$

Thus multiplication of an $n \times m$ matrix by an $m \times r$ matrix yields an $n \times r$ matrix. It should be noted that, in general, matrix multiplication is not commutative

$$\mathbf{AB} \neq \mathbf{BA} \qquad \text{or} \qquad \sum_{j=1}^{n} a_{ij}b_{jk} \neq \sum_{j=1}^{n} b_{ij}a_{jk}$$

For example

$$\mathbf{AB} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}$$

and

$$\mathbf{BA} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} b_{11}a_{11} + b_{12}a_{21} & b_{11}a_{12} + b_{12}a_{22} \\ b_{21}a_{11} + b_{22}a_{21} & b_{21}a_{12} + b_{22}a_{22} \end{bmatrix}$$

Thus, in general $\mathbf{AB} \neq \mathbf{BA}$. Hence, the order of multiplication is significant and must be preserved. If $\mathbf{AB} = \mathbf{BA}$, we call the matrices **A** and **B** *commute*. In the preceding matrices **A** and **B** if, for example, $a_{12} = a_{21} = b_{12} = b_{21} = 0$, then **A** and **B** commute.

For $n \times n$ diagonal matrices **A** and **B**

$$\mathbf{AB} = (a_{ij}\,\delta_{ij})(b_{ij}\,\delta_{ij}) = \begin{bmatrix} a_{11}b_{11} & & & 0 \\ & a_{22}b_{22} & & \\ & & \ddots & \\ 0 & & & a_{nn}b_{nn} \end{bmatrix}$$

If **A**, **B**, and **C** are $n \times m$ matrix, $m \times r$ matrix, and $r \times p$ matrix, respectively, then the following associativity law holds true: $(\mathbf{AB})\,\mathbf{C} = \mathbf{A}\,(\mathbf{BC})$. This may be proved as follows:

$$(i, k)\text{th element of } \mathbf{AB} = \sum_{j=1}^{m} a_{ij}b_{jk}$$

$$(j, h)\text{th element of } \mathbf{BC} = \sum_{k=1}^{r} b_{jk}c_{kh}$$

$$\begin{aligned}(i, h)\text{th element of } (\mathbf{AB})\mathbf{C} &= \sum_{k=1}^{r}\left(\sum_{j=1}^{m} a_{ij}b_{jk}\right)c_{kh} = \sum_{j=1}^{m}\sum_{k=1}^{r}(a_{ij}\,b_{jk})c_{kh} \\ &= \sum_{j=1}^{m}\sum_{k=1}^{r} a_{ij}(b_{jk}c_{kh}) = \sum_{j=1}^{m} a_{ij}\left(\sum_{k=1}^{r} b_{jk}c_{kh}\right) \\ &= (i, h)\text{th element of } \mathbf{A}(\mathbf{BC})\end{aligned}$$

Since associativity of multiplication of matrices holds true, we have

$$\mathbf{ABCD} = (\mathbf{AB})\,(\mathbf{CD}) = \mathbf{A}(\mathbf{BCD}) = (\mathbf{ABC})\mathbf{D}$$

$$\mathbf{A}^{m+n} = \mathbf{A}^{m}\mathbf{A}^{n} \qquad (m, n = 1, 2, 3, \ldots)$$

If **A** and **B** are $n \times m$ matrices and **C** and **D** are $m \times r$ matrices, then the following distributivity law holds true:

$$(\mathbf{A} + \mathbf{B})\,(\mathbf{C} + \mathbf{D}) = \mathbf{AC} + \mathbf{AD} + \mathbf{BC} + \mathbf{BD}$$

This can be proved by comparing the (i, j)th element of $(\mathbf{A} + \mathbf{B})(\mathbf{C} + \mathbf{D})$ and the (i, j)th element of $(\mathbf{AC} + \mathbf{AD} + \mathbf{BC} + \mathbf{BD})$.

Partitioning of matrix A matrix consisting of $q \times p$ $(p < m, q < n)$ array of elements obtained from an $n \times m$ matrix **A** by deleting any $n - q$ rows and any $m - p$ columns of **A** is called a *submatrix* of **A**. It is sometimes useful to partition **A** into submatrices and regard them as its elements. Partitioning of **A** into submatrices may be made in many different ways. For example, **A** may be partitioned as follows (the partitioning is denoted by the vertical and horizontal dotted lines):

$$\mathbf{A} = \left[\begin{array}{ccc:ccc} a_{11} & \cdots & a_{1p} & a_{1\,p+1} & \cdots & a_{1m} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{q1} & \cdots & a_{qp} & a_{q\,p+1} & \cdots & a_{qm} \\ \hdashline a_{q+1\,1} & \cdots & a_{q+1\,p} & a_{q+1\,p+1} & \cdots & a_{q+1\,m} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n1} & \cdots & a_{np} & a_{n\,p+1} & \cdots & a_{nm} \end{array}\right]$$

The elements of **A** contained in rectangles formed by dotted lines are the submatrix elements of **A**. By denoting the submatrices of this example by $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, $\mathbf{A}_{21}$, and $\mathbf{A}_{22}$, we obtain

$$\mathbf{A} = \left[\begin{array}{c:c} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \hdashline \mathbf{A}_{21} & \mathbf{A}_{22} \end{array}\right]$$

where

$$\mathbf{A}_{11} = \begin{bmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & & \vdots \\ a_{q1} & \cdots & a_{qp} \end{bmatrix} = q \times p \text{ matrix}$$

$$\mathbf{A}_{12} = \begin{bmatrix} a_{1\,p+1} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{q\,p+1} & \cdots & a_{qm} \end{bmatrix} = q \times (m - p) \text{ matrix}$$

$$\mathbf{A}_{21} = \begin{bmatrix} a_{q+1\,1} & \cdots & a_{q+1\,p} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{np} \end{bmatrix} = (n - q) \times p \text{ matrix}$$

$$\mathbf{A}_{22} = \begin{bmatrix} a_{q+1\,p+1} & \cdots & a_{q+1\,m} \\ \vdots & & \vdots \\ a_{n\,p+1} & \cdots & a_{nm} \end{bmatrix} = (n - q) \times (m - p) \text{ matrix}$$

If a matrix $\mathbf{B}$ is partitioned so that we can carry out submatrix multiplication with $\mathbf{A}$, for example,

$$\mathbf{B} = \left[\begin{array}{c|c} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \hline \mathbf{B}_{21} & \mathbf{B}_{22} \end{array}\right]$$

where

$$\mathbf{B}_{11} = \begin{bmatrix} b_{11} & \cdots & b_{1s} \\ \vdots & & \vdots \\ b_{p1} & \cdots & b_{ps} \end{bmatrix} = p \times s \text{ matrix}$$

$$\mathbf{B}_{12} = \begin{bmatrix} b_{1\,s+1} & \cdots & b_{1r} \\ \vdots & & \vdots \\ b_{p\,s+1} & \cdots & b_{pr} \end{bmatrix} = p \times (r-s) \text{ matrix}$$

$$\mathbf{B}_{21} = \begin{bmatrix} b_{p+1\,1} & \cdots & b_{p+1\,s} \\ \vdots & & \vdots \\ b_{m\,1} & \cdots & b_{m\,s} \end{bmatrix} = (m-p) \times s \text{ matrix}$$

$$\mathbf{B}_{22} = \begin{bmatrix} b_{p+1\,s+1} & \cdots & b_{p+1\,r} \\ \vdots & & \vdots \\ b_{m\,s+1} & \cdots & b_{mr} \end{bmatrix} = (m-p) \times (r-s) \text{ matrix}$$

then

$$\mathbf{AB} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{bmatrix}$$

This equivalence may be seen if we examine the (i, k)th element of $\mathbf{AB}$. Assume that $1 \leqslant i \leqslant q$, $1 \leqslant k \leqslant s$. Then

$$(i,k)\text{th element of } \mathbf{AB} = \sum_{j=1}^{m} a_{ij}b_{jk} = \sum_{j=1}^{p} a_{ij}b_{jk} + \sum_{j=1}^{m-p} a_{i\,p+j}b_{p+j\,k}$$

$$= \sum_{j=1}^{p} [(i,j)\text{th element of } \mathbf{A}_{11}][(j,k)\text{th element of } \mathbf{B}_{11}]$$

$$+ \sum_{j=1}^{m-p} [(i,j)\text{th element of } \mathbf{A}_{12}][(j,k)\text{th element of } \mathbf{B}_{21}]$$

$$= (i,k)\text{th element of } (\mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21})$$

Hence, we have the following relationships:

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}\begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{0} & \mathbf{B}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{B}_{11} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{0} & \mathbf{A}_{22}\mathbf{B}_{22} \end{bmatrix}$$

and

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}\begin{bmatrix} \mathbf{B}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{B}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22}\mathbf{B}_{22} \end{bmatrix}$$

Notice that if $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are nonsingular, then

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{0} & \mathbf{A}_{22}^{-1} \end{bmatrix}$$

and

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{A}_{22}^{-1} \end{bmatrix}$$

Partitioning is useful in additions and multiplications of matrices because the computing labor can be lessened considerably. To demonstrate this, consider multiplication of two matrices. If two matrices $\mathbf{A}$ and $\mathbf{B}$ are given by

$$\mathbf{A} = \left[\begin{array}{cc|cc} 1 & 2 & 2 & 4 \\ -3 & 1 & 1 & -1 \end{array}\right] = [\mathbf{A}_{11} \mid \mathbf{A}_{12}]$$

$$\mathbf{B} = \left[\begin{array}{cc} 3 & 0 \\ 1 & 0 \\ \hline 2 & 1 \\ 0 & 1 \end{array}\right] = \begin{bmatrix} \mathbf{B}_{11} \\ \hline \mathbf{B}_{21} \end{bmatrix}$$

then $\mathbf{AB}$ is obtained as follows:

$$\mathbf{AB} = [\mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21}]$$

$$= \begin{bmatrix} 1 & 2 \\ -3 & 1 \end{bmatrix}\begin{bmatrix} 3 & 0 \\ 1 & 0 \end{bmatrix} + \begin{bmatrix} 2 & 4 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 9 & 6 \\ -6 & 0 \end{bmatrix}$$

Remarks on cancellation of matrices Cancellation of matrices is not valid in matrix algebra. Consider the product of two singular matrices $\mathbf{A}$ and $\mathbf{B}$. Take, for example,

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 6 & 3 \end{bmatrix} \neq \mathbf{0}$$

$$\mathbf{B} = \begin{bmatrix} 1 & -2 \\ -2 & 4 \end{bmatrix} \neq \mathbf{0}$$

Then,

$$\mathbf{AB} = \begin{bmatrix} 2 & 1 \\ 6 & 3 \end{bmatrix}\begin{bmatrix} 1 & -2 \\ -2 & 4 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \mathbf{0}$$

Clearly $\mathbf{AB} = \mathbf{0}$ implies neither $\mathbf{A} = \mathbf{0}$ nor $\mathbf{B} = \mathbf{0}$. In fact $\mathbf{AB} = \mathbf{0}$ implies one of the following three:

(a) $\mathbf{A} = \mathbf{0}$
(b) $\mathbf{B} = \mathbf{0}$
(c) Both $\mathbf{A}$ and $\mathbf{B}$ are singular

We can easily prove that if both $\mathbf{A}$ and $\mathbf{B}$ are nonzero matrices and

$\mathbf{AB} = \mathbf{0}$, then both $\mathbf{A}$ and $\mathbf{B}$ are singular. Assume that if $\mathbf{A}$ and $\mathbf{B}$ are not singular, then $|\mathbf{A}| \neq 0$ and therefore $\mathbf{A}^{-1}$ exists. Then we obtain

$$\mathbf{A}^{-1}\mathbf{AB} = \mathbf{B} = \mathbf{0}$$

which contradicts the assumption that $\mathbf{B}$ is nonzero. Thus we conclude that both A and B must be singular if $\mathbf{A} \neq \mathbf{0}$ and $\mathbf{B} \neq \mathbf{0}$.

Similarly, notice that if $\mathbf{A}$ is singular, then neither $\mathbf{AB} = \mathbf{AC}$ nor $\mathbf{BA} = \mathbf{CA}$ implies $\mathbf{B} = \mathbf{C}$. If, however, $\mathbf{A}$ is a nonsingular matrix, then $\mathbf{AB} = \mathbf{AC}$ implies $\mathbf{B} = \mathbf{C}$ and $\mathbf{BA} = \mathbf{CA}$ also implies $\mathbf{B} = \mathbf{C}$.

Derivative or integral of matrix The time *derivative* of a matrix $\mathbf{A}(t)$ is defined by the matrix whose (i, j)th element is the time derivative of the (i, j)th element of the original matrix, provided that all the elements $a_{ij}(t)$ have derivatives with respect to t. Namely,

$$\frac{d}{dt}\mathbf{A}(t) = \left(\frac{d}{dt}a_{ij}(t)\right)$$

Similarly, the time *integral* of a matrix $\mathbf{A}(t)$ is defined by the matrix whose (i, j)th element is the time integral of the (i, j)th element of the original matrix, or

$$\int \mathbf{A}(t)\,dt = \left(\int a_{ij}(t)\,dt\right)$$

provided that the $a_{ij}(t)$ are integrable as functions of t.

Differentiation of matrix If elements of matrices $\mathbf{A}$ and $\mathbf{B}$ are functions of time, then

$$\frac{d}{dt}(\mathbf{AB}) = \frac{d\mathbf{A}}{dt}\mathbf{B} + \mathbf{A}\frac{d\mathbf{B}}{dt}$$

$$\frac{d}{dt}(\mathbf{A} + \mathbf{B}) = \frac{d\mathbf{A}}{dt} + \frac{d\mathbf{B}}{dt}$$

Here again multiplication of $\mathbf{A}$ and $d\mathbf{B}/dt$ is, in general, not commutative. If α is a scalar and is a function of time, then

$$\frac{d}{dt}(\mathbf{A}\alpha) = \frac{d\mathbf{A}}{dt}\alpha + \mathbf{A}\frac{d\alpha}{dt}$$

The time derivative of $\mathbf{A}^{-1}$ is given by

$$\frac{d}{dt}\mathbf{A}^{-1} = -\mathbf{A}^{-1}\frac{d\mathbf{A}}{dt}\mathbf{A}^{-1}$$

This may be obtained by differentiating $\mathbf{AA}^{-1}$ with respect to t. Since

$$\frac{d}{dt}\mathbf{AA}^{-1} = \frac{d\mathbf{A}}{dt}\mathbf{A}^{-1} + \mathbf{A}\frac{d\mathbf{A}^{-1}}{dt}$$

$$\frac{d}{dt}\mathbf{AA}^{-1} = \frac{d}{dt}\mathbf{I} = \mathbf{0}$$

we obtain

$$\mathbf{A}\frac{d\mathbf{A}^{-1}}{dt} = -\frac{d\mathbf{A}}{dt}\mathbf{A}^{-1} \qquad \text{or} \qquad \mathbf{A}^{-1}\mathbf{A}\frac{d\mathbf{A}^{-1}}{dt} = -\mathbf{A}^{-1}\frac{d\mathbf{A}}{dt}\mathbf{A}^{-1}$$

which is the desired result.

2-3 VECTORS AND VECTOR SPACES

This section presents definitions and brief discussions of vectors and vector spaces.

Vector Let F denote the real number field or the complex number field. The elements of F are termed *scalars*. An n-dimensional vector $\mathbf{x}$, called *n vector*, over a field F is an ordered set of n numbers which belong to F. We denote it in the form:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$

The n numbers $x_1, x_2, \ldots, x_n$ are called *elements* or *components* of $\mathbf{x}$. Vectors over the field of real numbers are called *real* vectors; vectors over the field of complex numbers are called *complex* vectors. The vector previously denoted is called a *column* vector.[1] The column vector may be considered as an $n \times 1$ matrix. The transpose of a real column vector $\mathbf{x}$ is a real row vector denoted by

$$\mathbf{x}' = [x_1 \;\; x_2 \;\; \cdots \;\; x_n]$$

The conjugate transpose of a complex column vector is a complex row vector denoted by

$$\mathbf{x}^* = [\bar{x}_1 \;\; \bar{x}_2 \;\; \cdots \;\; \bar{x}_n]$$

Note that the conjugate transpose of a real column vector is simply a real row vector.

One-dimensional vectors are *scalars*. (Scalars are either real numbers or complex numbers.)

[1] In this book all vectors imply column vectors. Any row vector is written as the transpose or conjugate transpose of a column vector.

If $\mathbf{A}$ is an $n \times n$ matrix and $\mathbf{x}$ is an n vector, then the product $\mathbf{Ax}$ is an $n \times 1$ matrix or column vector.

Two vectors $\mathbf{x}$ and $\mathbf{y}$ are called *equal*, if and only if their corresponding components are equal, or $x_i = y_i$ $(i = 1, 2, \ldots, n)$.

Null vector The vector whose components are all zero is called the *null vector*. It is denoted by $\mathbf{0}$.

Vector space A *vector space* V over a field F is defined as a set of all vectors over F for which the operations of addition and multiplication by scalars are defined and satisfy the following rules:

1. For each pair of vectors $\mathbf{x}$ and $\mathbf{y}$ in V, the sum of $\mathbf{x}$ and $\mathbf{y}$, denoted by $\mathbf{x} + \mathbf{y}$, belongs to V. Addition of vectors is commutative and associative, namely

$$\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$$
$$(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$$

2. For each pair of vectors $\mathbf{x}$ and $\mathbf{y}$ in V there exists a unique vector $\mathbf{z}$ in V such that $\mathbf{x} + \mathbf{z} = \mathbf{y}$. In particular, there exists a unique vector denoted by $\mathbf{0}$ such that $\mathbf{x} + \mathbf{0} = \mathbf{x}$ for all $\mathbf{x}$ in V. To every vector $\mathbf{x}$ in V, there corresponds a unique vector $-\mathbf{x}$ such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$.
3. For each vector $\mathbf{x}$ in V and each scalar c in F the product $c\mathbf{x}$ belongs to V. Multiplication of vectors by scalars is associative and distributive, namely,

$$(c_1c_2)\mathbf{x} = c_1(c_2\mathbf{x})$$
$$c(\mathbf{x} + \mathbf{y}) = c\mathbf{x} + c\mathbf{y}$$
$$(c_1 + c_2)\mathbf{x} = c_1\mathbf{x} + c_2\mathbf{x}$$

The product of the identity matrix $\mathbf{I}$ and a vector $\mathbf{x}$ is $\mathbf{x}$ itself, or

$$\mathbf{Ix} = \mathbf{x}$$

The foregoing definition implies that a vector space is a set of vectors closed with respect to vector addition and multiplication of vectors by scalars.

If F is the field of real numbers, V is called a *real* vector space. Similarly, if F is the field of complex numbers, V is called a *complex* vector space. In this book all vectors and scalars treated are assumed to be elements of a complex vector space unless the contrary is stated.

Linear dependence and independence Vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are said to be *linearly independent* if

$$c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_n\mathbf{x}_n = \mathbf{0}$$

where $c_1, c_2, \ldots, c_n$ are constants, implies that $c_1 = c_2 = \cdots = c_n = 0$. Conversely, vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are said to be *linearly dependent* if and only if $\mathbf{x}_i$ can be expressed as a linear combination of $\mathbf{x}_j$ $(j = 1, 2, \ldots, n;\ j \neq i)$.

Theorem If vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are linearly independent and vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \mathbf{x}_{n+1}$ are linearly dependent, then $\mathbf{x}_{n+1}$ can be expressed as a unique linear combination of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$.

Proof Since vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n+1}$ are linearly dependent,

$$c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_n\mathbf{x}_n + c_{n+1}\mathbf{x}_{n+1} = \mathbf{0}$$

where not all c_i are zero. Clearly $c_{n+1} \neq 0$. (Otherwise, the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ become linearly dependent which contradicts the assumption that they are linearly independent.) Hence,

$$\mathbf{x}_{n+1} = -\frac{c_1}{c_{n+1}}\mathbf{x}_1 - \frac{c_2}{c_{n+1}}\mathbf{x}_2 - \cdots - \frac{c_n}{c_{n+1}}\mathbf{x}_n$$

Thus, $\mathbf{x}_{n+1}$ can be expressed as a linear combination of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. The uniqueness of this expression can be proved as follows: Suppose that there are two expressions

$$\mathbf{x}_{n+1} = \sum_{i=1}^{n} c_i'\mathbf{x}_i = \sum_{i=1}^{n} c_i''\mathbf{x}_i$$

then clearly

$$\sum_{i=1}^{n} (c_i' - c_i'')\mathbf{x}_i = \mathbf{0}$$

Since vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are linearly independent,

$$c_i' - c_i'' = 0$$

Namely,

$$c_i' = c_i'' \qquad (1 \leqslant i \leqslant n)$$

This proves the uniqueness of the expression for $\mathbf{x}_{n+1}$.

It can be proved that the necessary and sufficient conditions for n vectors $\mathbf{x}_i$ $(i = 1, 2, \ldots, m)$ to be linearly independent are that

1. $m \leqslant n$
2. There exists at least one nonzero determinant made of m columns of the $n \times m$ matrix whose m columns consist of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$

Hence for n vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, the necessary and sufficient condition for linear independence is

$$|\mathbf{A}| \neq 0$$

where $\mathbf{A}$ is the $n \times n$ matrix whose ith column is made of the components of $\mathbf{x}_i$ $(i = 1, 2, \ldots, n)$.

Basis A set of vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ of a vector space V is called a *basis* of V if these n vectors are linearly independent and every vector $\mathbf{y}$ of the vector space V can be uniquely expressed as a linear combination of the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, namely,

$$\mathbf{y} = \sum_{i=1}^{n} a_i \mathbf{x}_i$$

where the numbers a_i are the coordinates of $\mathbf{y}$ relative to the basis $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. We shall denote the basis $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ by the symbol

$$\mathscr{B} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$$

Finite dimensional vector space A vector space V is called a *finite dimensional* vector space if the number of elements in its basis is finite. The number of elements in a basis of a finite dimensional vector space V is called the *dimension* of V. In this book we denote n-dimensional vector space by V^n or simply by V.

For a given vector space V a basis is not unique. In an n-dimensional vector space any basis has n elements. Conversely, any set of n linearly independent n vectors is a basis of an n-dimensional vector space. For example, the following vectors

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{x}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

constitute a basis of a three-dimensional vector space. The following three vectors, however,

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{x}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

do not constitute a basis of a three-dimensional vector space, since they are not linearly independent.

If any vector $\mathbf{y}$ of an n-dimensional vector space V can be expressed as a linear combination of a set of vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, then the set of vectors is said to span the n-dimensional vector space V. Consider the following set of n vectors:

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \quad \ldots, \quad \mathbf{x}_n = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ 1 \end{bmatrix}$$

Since any vector $\mathbf{y}$ where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}$$

can be written as

$$\mathbf{y} = y_1\mathbf{x}_1 + y_2\mathbf{x}_2 + \cdots + y_n\mathbf{x}_n$$

the set of vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ spans n-dimensional vector space V.

Subspace Any nonempty set S of vectors of V is called a *subspace* of V if it is closed with respect to vector addition and multiplication of vectors by scalars. Namely, if any pair of vectors $\mathbf{x}$ and $\mathbf{y}$ belongs to S, then $\mathbf{x} + \mathbf{y}$ also belongs to S; and if any vector $\mathbf{x}$ belongs to S and c is any scalar, then $c\mathbf{x}$ also belongs to S. Hence a subspace always contains $0\mathbf{x} = \mathbf{0}$. A subspace of a vector space is itself a vector space.

For any set of vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ the set of all their linear combinations constitutes a subspace, and the subspace is spanned by them. If a set of vectors spans a subspace and if in addition these vectors are linearly independent, then the set of vectors is called a *basis of the subspace*. A basis of a subspace can always be found from a set of vectors which spans the subspace by choosing a maximum number of linearly independent vectors from the set. Notice that the number of vectors in every basis of a subspace is the same and is equal to the dimension of the subspace.

Linear transformation Consider a linear transformation

$$\mathbf{z} = \mathbf{A}\mathbf{x}$$

where

$\mathbf{x} = m$ vector

$\mathbf{z} = n$ vector

$\mathbf{A} = n \times m$ matrix

The matrix $\mathbf{A}$ transforms a vector $\mathbf{x}$ in an m-dimensional vector space into a vector $\mathbf{z}$ in an n-dimensional vector space. The important problem in connection with linear transformations is the problem of inversion, the question of the existence of a solution to a system of linear equations. We shall discuss this matter in Chapter 7.

Inner product Any rule, which assigns to each pair of vectors $\mathbf{x}$ and $\mathbf{y}$ in a vector space a scalar quantity $(\mathbf{x}, \mathbf{y})$, is called an *inner* product or

scalar product in the vector space provided that the following four axioms are satisfied:

1. $(\mathbf{y}, \mathbf{x}) = \overline{(\mathbf{x}, \mathbf{y})}$
   where the bar denotes the conjugate of a complex number.
2. $(c\mathbf{x}, \mathbf{y}) = \bar{c}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \bar{c}\mathbf{y})$
   where c is a complex number.
3. $(\mathbf{x} + \mathbf{y}, \mathbf{z} + \mathbf{w}) = (\mathbf{x}, \mathbf{z}) + (\mathbf{y}, \mathbf{z}) + (\mathbf{x}, \mathbf{w}) + (\mathbf{y}, \mathbf{w})$
4. $(\mathbf{x}, \mathbf{x}) > 0$ for $\mathbf{x} \neq \mathbf{0}$

In any finite dimensional vector space, there are many different definitions of inner product, all having the properties required by the definition.

In this book, unless the contrary is stated, we shall adopt the following definition of the *inner product:* The inner product of a pair of n vectors $\mathbf{x}$ and $\mathbf{y}$ in a vector space V is given by

$$(\mathbf{x}, \mathbf{y}) = \bar{x}_1 y_1 + \bar{x}_2 y_2 + \cdots + \bar{x}_n y_n = \sum_{i=1}^{n} \bar{x}_i y_i$$
$$= \text{complex number}$$

where the $\bar{x}_i$ are the complex conjugates of the x_i. This definition clearly satisfies the four axioms. The inner product can then be expressed as $(\mathbf{x}, \mathbf{y}) = \mathbf{x}^*\mathbf{y}$, where $\mathbf{x}^*$ denotes the conjugate transpose of $\mathbf{x}$. Also,

$$(\mathbf{x}, \mathbf{y}) = \overline{(\mathbf{y}, \mathbf{x})} = \overline{\mathbf{y}^* \mathbf{x}} = \mathbf{y}'\bar{\mathbf{x}} = \mathbf{x}^*\mathbf{y}$$

where $\mathbf{y}'$ is the transpose of $\mathbf{y}$.

The inner product of two n vectors $\mathbf{x}$ and $\mathbf{y}$ with real components is therefore given by

$$(\mathbf{x}, \mathbf{y}) = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = \sum_{i=1}^{n} x_i y_i$$

In this case clearly,

$$(\mathbf{x}, \mathbf{y}) = \mathbf{x}'\mathbf{y} = \mathbf{y}'\mathbf{x} \qquad \text{(for real vectors } \mathbf{x} \text{ and } \mathbf{y})$$

The real or complex vector $\mathbf{x}$ is said to be *normalized* if $(\mathbf{x}, \mathbf{x}) = 1$.

We should note that for an n vector $\mathbf{x}$, $\mathbf{x}^*\mathbf{x}$ is a nonnegative scalar but $\mathbf{x}\mathbf{x}^*$ is an $n \times n$ matrix. Namely,

$$\mathbf{x}^*\mathbf{x} = \bar{x}_1 x_1 + \bar{x}_2 x_2 + \cdots + \bar{x}_n x_n$$
$$= |x_1|^2 + |x_2|^2 + \cdots + |x_n|^2$$

and

$$\mathbf{x}\mathbf{x}^* = \begin{bmatrix} x_1\bar{x}_1 & x_1\bar{x}_2 & \cdots & x_1\bar{x}_n \\ x_2\bar{x}_1 & x_2\bar{x}_2 & \cdots & x_2\bar{x}_n \\ \vdots & \vdots & & \vdots \\ x_n\bar{x}_1 & x_n\bar{x}_2 & \cdots & x_n\bar{x}_n \end{bmatrix}$$

Notice that, for a complex $n \times n$ matrix $\mathbf{A}$ and complex n vectors $\mathbf{x}$ and $\mathbf{y}$, the inner product of $\mathbf{Ax}$ and $\mathbf{y}$ has the property that $(\mathbf{Ax}, \mathbf{y}) = (\mathbf{x}, \mathbf{A}^*\mathbf{y})$, where $\mathbf{A}^*$ is the conjugate transpose of $\mathbf{A}$.

For a real $n \times n$ matrix $\mathbf{A}$ and real n vectors $\mathbf{x}$ and $\mathbf{y}$, the inner product of $\mathbf{Ax}$ and $\mathbf{y}$ has the property that $(\mathbf{Ax}, \mathbf{y}) = (\mathbf{x}, \mathbf{A}'\mathbf{y})$, where $\mathbf{A}'$ is the transpose of $\mathbf{A}$.

Unitary transformation If $\mathbf{A}$ is a unitary matrix ($\mathbf{A}^{-1} = \mathbf{A}^*$), then the inner product $(\mathbf{x}, \mathbf{x})$ is invariant under the linear transformation $\mathbf{x} = \mathbf{Ay}$, because

$$(\mathbf{x}, \mathbf{x}) = (\mathbf{Ay}, \mathbf{Ay}) = (\mathbf{y}, \mathbf{A}^*\mathbf{Ay}) = (\mathbf{y}, \mathbf{A}^{-1}\mathbf{Ay}) = (\mathbf{y}, \mathbf{y})$$

Such a transformation $\mathbf{x} = \mathbf{Ay}$, where $\mathbf{A}$ is a unitary matrix, that transforms $\sum_{i=1}^{n} \bar{x}_i x_i$ into $\sum_{i=1}^{n} \bar{y}_i y_i$ is called a *unitary* transformation.

Orthogonal transformation If $\mathbf{A}$ is an orthogonal matrix ($\mathbf{A}^{-1} = \mathbf{A}'$), then the inner product $(\mathbf{x}, \mathbf{x})$ is invariant under the linear transformation $\mathbf{x} = \mathbf{Ay}$, because

$$(\mathbf{x}, \mathbf{x}) = (\mathbf{Ay}, \mathbf{Ay}) = (\mathbf{y}, \mathbf{A}'\mathbf{Ay}) = (\mathbf{y}, \mathbf{A}^{-1}\mathbf{Ay}) = (\mathbf{y}, \mathbf{y})$$

Such a transformation $\mathbf{x} = \mathbf{Ay}$ that transforms $\sum_{i=1}^{n} x_i^2$ into $\sum_{i=1}^{n} y_i^2$ is called an *orthogonal* transformation.

Norms of vector Once we define the inner product we can use this inner product to define norms of a vector $\mathbf{x}$. The concept of a norm is somewhat similar to that of the absolute value. A norm is a function which assigns to every vector $\mathbf{x}$ in a given vector space a real number denoted by $\|\mathbf{x}\|$ such that

1. $\|\mathbf{x}\| > 0$ for $\mathbf{x} \neq \mathbf{0}$
2. $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$
3. $\|a\mathbf{x}\| = |a|\,\|\mathbf{x}\|$
   where a is a scalar and $|a|$ is the absolute value of a.
4. $\|\mathbf{x} + \mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$ for all $\mathbf{x}$ and $\mathbf{y}$
5. $|(\mathbf{x}, \mathbf{y})| \leqslant \|\mathbf{x}\|\,\|\mathbf{y}\|$ (Schwarz inequality)

Several different definitions of norms are commonly used in the literature. In this book, however, we shall employ the following definition: A *norm* of a vector is defined as the nonnegative square root of $(\mathbf{x}, \mathbf{x})$, or

$$\|\mathbf{x}\| = (\mathbf{x}, \mathbf{x})^{1/2} = [\mathbf{x}^*\mathbf{x}]^{1/2} = \sqrt{|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2}$$

If $\mathbf{x}$ is a real vector, the quantity $\|\mathbf{x}\|^2$ can be interpreted geometrically as the square of the distance from the origin to the point represented by the vector $\mathbf{x}$.

The five properties of norms listed earlier may be obvious, except perhaps the last two inequalities. We can prove these two inequalities as follows: From definitions of the inner product and norm, we have

$$\begin{aligned}\|\lambda\mathbf{x}+\mathbf{y}\|^2 &= (\lambda\mathbf{x}+\mathbf{y},\lambda\mathbf{x}+\mathbf{y}) = (\lambda\mathbf{x},\lambda\mathbf{x})+(\mathbf{y},\lambda\mathbf{x})+(\lambda\mathbf{x},\mathbf{y})+(\mathbf{y},\mathbf{y})\\ &= \bar{\lambda}\lambda\|\mathbf{x}\|^2+\lambda(\mathbf{y},\mathbf{x})+\bar{\lambda}(\mathbf{x},\mathbf{y})+\|\mathbf{y}\|^2\\ &= \bar{\lambda}[\lambda\|\mathbf{x}\|^2+(\mathbf{x},\mathbf{y})]+\lambda(\overline{\mathbf{x},\mathbf{y}})+\|\mathbf{y}\|^2 \geqslant 0\end{aligned}$$

If we choose

$$\lambda = -\frac{(\mathbf{x},\mathbf{y})}{\|\mathbf{x}\|^2} \qquad \text{for } \mathbf{x}\neq\mathbf{0}$$

then

$$\lambda(\overline{\mathbf{x},\mathbf{y}})+\|\mathbf{y}\|^2 = -\frac{(\mathbf{x},\mathbf{y})(\overline{\mathbf{x},\mathbf{y}})}{\|\mathbf{x}\|^2}+\|\mathbf{y}\|^2 \geqslant 0$$

and

$$\|\mathbf{x}\|^2\|\mathbf{y}\|^2 \geqslant (\mathbf{x},\mathbf{y})(\overline{\mathbf{x},\mathbf{y}}) = |(\mathbf{x},\mathbf{y})|^2 \qquad \text{for } \mathbf{x}\neq\mathbf{0}$$

For $\mathbf{x}=\mathbf{0}$, clearly,

$$\|\mathbf{x}\|^2\|\mathbf{y}\|^2 = |(\mathbf{x},\mathbf{y})|^2$$

Therefore, we obtain the Schwarz inequality,

$$|(\mathbf{x},\mathbf{y})| \leqslant \|\mathbf{x}\|\|\mathbf{y}\|$$

By the use of the Schwarz inequality we obtain the following inequality

$$\|\mathbf{x}+\mathbf{y}\| \leqslant \|\mathbf{x}\|+\|\mathbf{y}\|$$

This can be proved easily, since

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|^2 &= (\mathbf{x}+\mathbf{y},\mathbf{x}+\mathbf{y})\\ &= (\mathbf{x},\mathbf{x})+(\mathbf{x},\mathbf{y})+(\mathbf{y},\mathbf{x})+(\mathbf{y},\mathbf{y})\\ &= \|\mathbf{x}\|^2+(\mathbf{x},\mathbf{y})+(\overline{\mathbf{x},\mathbf{y}})+\|\mathbf{y}\|^2\\ &= \|\mathbf{x}\|^2+\|\mathbf{y}\|^2+2\mathrm{Re}(\mathbf{x},\mathbf{y})\\ &\leqslant \|\mathbf{x}\|^2+\|\mathbf{y}\|^2+2|(\mathbf{x},\mathbf{y})|\\ &\leqslant \|\mathbf{x}\|^2+\|\mathbf{y}\|^2+2\|\mathbf{x}\|\|\mathbf{y}\|\\ &= (\|\mathbf{x}\|+\|\mathbf{y}\|)^2\end{aligned}$$

Three other definitions of norms commonly used in the literature (2–5)[1] are given:

(1) A norm $\|\mathbf{x}\|$ may be defined by

$$\begin{aligned}\|\mathbf{x}\| &= [(\mathbf{Tx})^*(\mathbf{Tx})]^{1/2} = [\mathbf{x}^*\mathbf{T}^*\mathbf{Tx}]^{1/2} = [\mathbf{x}^*\mathbf{Qx}]^{1/2}\\ &= \left[\sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij}\bar{x}_i x_j\right]^{1/2} \geqslant 0\end{aligned}$$

[1]Numbers in parentheses refer to the list of references at the end of this chapter.

The matrix $\mathbf{Q}=\mathbf{T}^*\mathbf{T}$ is Hermitian (since $\mathbf{Q}^*=\mathbf{T}^*\mathbf{T}=\mathbf{Q}$). The norm $\|\mathbf{x}\|=[\mathbf{x}^*\mathbf{Qx}]^{1/2}$ is a generalized form of $[\mathbf{x}^*\mathbf{x}]^{1/2}$ which can be written as $[\mathbf{x}^*\mathbf{I}\mathbf{x}]^{1/2}$.

(2) A *norm* may be defined as the sum of the magnitudes of all the components x_i, or

$$\|\mathbf{x}\| = \sum_{i=1}^{n}|x_i|$$

(3) A *norm* may be defined as the maximum of the magnitudes of all the components x_i, or

$$\|\mathbf{x}\| = \max_i\{|x_i|\}$$

Among many definitions of norms, the norm $[\mathbf{x}^*\mathbf{x}]^{1/2}$ is most commonly used in explicit calculations. The concept of norms is useful to prove convergence of series. It can be shown that the various norms just defined are equivalent. To prove convergence of series we may use any of the foregoing norms.

Orthogonality of vectors If the inner product of two vectors $\mathbf{x}$ and $\mathbf{y}$ is zero, or $(\mathbf{x},\mathbf{y})=0$, then vectors $\mathbf{x}$ and $\mathbf{y}$ are said to be *orthogonal* to each other. For example, vectors

$$\mathbf{x}_1=\begin{bmatrix}1\\1\\0\end{bmatrix}, \qquad \mathbf{x}_2=\begin{bmatrix}0\\0\\1\end{bmatrix}, \qquad \mathbf{x}_3=\begin{bmatrix}1\\-1\\0\end{bmatrix}$$

are orthogonal in pairs and thus form an orthogonal set. In an n-dimensional vector space, vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ defined by

$$\mathbf{x}_1=\begin{bmatrix}1\\0\\0\\\cdot\\\cdot\\\cdot\\0\end{bmatrix}, \qquad \mathbf{x}_2=\begin{bmatrix}0\\1\\0\\\cdot\\\cdot\\\cdot\\0\end{bmatrix}, \quad \ldots, \quad \mathbf{x}_n=\begin{bmatrix}0\\0\\0\\\cdot\\\cdot\\\cdot\\1\end{bmatrix}$$

satisfy the conditions $(\mathbf{x}_i,\mathbf{x}_j)=\delta_{ij}$, or

$$\begin{aligned}(\mathbf{x}_i,\mathbf{x}_j) &= 0 \qquad (i\neq j)\\ (\mathbf{x}_i,\mathbf{x}_i) &= 1\end{aligned}$$

where $i,j=1,2,\ldots,n$. Such a set of vectors is called *orthonormal*, since the vectors are orthogonal to each other and each vector is normalized. A nonzero vector $\mathbf{x}$ can be normalized by dividing $\mathbf{x}$ by $\|\mathbf{x}\|$. The normalized

vector $\mathbf{x}/\|\mathbf{x}\|$ is a unit vector. Unit vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ form an orthonormal set if they are orthogonal in pairs.

Consider a unitary matrix $\mathbf{A}$. By partitioning $\mathbf{A}$ into column vectors $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n$, we have

$$\mathbf{A}^*\mathbf{A} = \begin{bmatrix} \mathbf{A}_1^* \\ \hline \mathbf{A}_2^* \\ \hline \vdots \\ \hline \mathbf{A}_n^* \end{bmatrix} [\mathbf{A}_1 \mid \mathbf{A}_2 \mid \cdots \mid \mathbf{A}_n]$$

$$= \begin{bmatrix} \mathbf{A}_1^*\mathbf{A}_1 & \mathbf{A}_1^*\mathbf{A}_2 & \cdots & \mathbf{A}_1^*\mathbf{A}_n \\ \mathbf{A}_2^*\mathbf{A}_1 & \mathbf{A}_2^*\mathbf{A}_2 & \cdots & \mathbf{A}_2^*\mathbf{A}_n \\ \vdots & \vdots & & \vdots \\ \mathbf{A}_n^*\mathbf{A}_1 & \mathbf{A}_n^*\mathbf{A}_2 & \cdots & \mathbf{A}_n^*\mathbf{A}_n \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

it follows that

$$\mathbf{A}_i^*\mathbf{A}_i = (\mathbf{A}_i, \mathbf{A}_i) = 1$$

$$\mathbf{A}_i^*\mathbf{A}_j = (\mathbf{A}_i, \mathbf{A}_j) = 0 \qquad (i \neq j)$$

Thus we see that the column vectors (or row vectors) of a unitary matrix $\mathbf{A}$ are orthonormal. The same is true for orthogonal matrices, since they are unitary.

Notice that a real (complex) matrix is orthogonal (unitary) if and only if its columns or rows form an orthonormal set of vectors. Notice also that if nonzero vectors form an orthogonal set, then they are linearly independent. If a vector $\mathbf{x}$ of an n-dimensional vector space V is orthogonal to all vectors of a basis of V, then $\mathbf{x} = \mathbf{0}$.

Dual space and dual basis The vector space V' consisting of all linear functionals on V is called the *dual* vector space to V. If a set of vectors $\mathscr{B} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is a basis of an n-dimensional vector space V, then there exists a uniquely determined basis $\mathscr{B}' = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$ in V' such that

$$(\mathbf{y}_i, \mathbf{x}_j) = \delta_{ij} \qquad (i, j = 1, 2, \ldots, n)$$

The basis $\mathscr{B}'$ is called the *dual basis* of $\mathscr{B}$. The dual vector space of an n-dimensional vector space is also n-dimensional. (Notice that the vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ are linearly independent and span the n-dimensional space V'. Hence they constitute a basis of V'.)

Any basis has a unique dual basis. The n components of $\bar{\mathbf{y}}_i$ $(i = 1, 2, \ldots, n)$, namely $\bar{y}_{1i}, \bar{y}_{2i}, \ldots, \bar{y}_{ni}$, can be determined by solving the following algebraic equations:

$$\begin{aligned} \bar{y}_{1i}x_{11} + \bar{y}_{2i}x_{21} + \cdots + \bar{y}_{ni}x_{n1} &= \delta_{i1} \\ \bar{y}_{1i}x_{12} + \bar{y}_{2i}x_{22} + \cdots + \bar{y}_{ni}x_{n2} &= \delta_{i2} \\ \cdots \qquad \cdots \qquad \cdots & \\ \bar{y}_{1i}x_{1n} + \bar{y}_{2i}x_{2n} + \cdots + \bar{y}_{ni}x_{nn} &= \delta_{in} \\ (i = 1, 2, \ldots, n) & \end{aligned}$$

Notice that in the foregoing n simultaneous equations in $\bar{y}_{1i}, \bar{y}_{2i}, \ldots, \bar{y}_{ni}$, the coefficient determinant is nonzero, or

$$\begin{vmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{nn} \end{vmatrix} \neq 0$$

since the rows of this determinant consist of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ which are linearly independent. Hence n components of $\bar{\mathbf{y}}_i$ can be determined uniquely. Thus the dual basis can be determined by the process of finding an inverse matrix and taking its conjugate transpose. This process is further explained in the following: Consider a matrix $\mathbf{A}$ consisting of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ which constitute a basis of an n-dimensional vector space.

$$\mathbf{A} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix}$$

Consider next a matrix $\mathbf{B}$ whose columns consist of $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ which constitute the dual basis of the basis $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$.

$$\mathbf{B} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nn} \end{bmatrix}$$

Then the product $\mathbf{B}^*\mathbf{A}$ becomes as follows:

$$\mathbf{B}^*\mathbf{A} = \left(\sum_{k=1}^{n} \bar{y}_{ki} x_{kj}\right) = (\delta_{ij}) = \mathbf{I}$$

Hence, we see that $\mathbf{B} = (\mathbf{A}^{-1})^*$.

For an orthonormal basis $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, namely, $(\mathbf{x}_i, \mathbf{x}_j) = \delta_{ij}$, the basis is itself its own dual basis.

Any vector $\mathbf{z}$ of an n-dimensional vector space V can be uniquely expressed relative to a basis $\mathscr{B} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ as follows:

$$\mathbf{z} = \sum_{i=1}^{n} (\mathbf{y}_i, \mathbf{z})\mathbf{x}_i$$

where $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ are the dual basis of a basis $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. This can be easily proved. Since $\mathbf{z}$ can be represented as

$$\mathbf{z} = \sum_{i=1}^{n} a_i \mathbf{x}_i$$

the inner product of $\mathbf{y}_i$ and $\mathbf{z}$ is given by

$$(\mathbf{y}_i, \mathbf{z}) = \left(\mathbf{y}_i, \sum_{i=1}^{n} a_i \mathbf{x}_i\right) = a_i$$

Hence,

$$\mathbf{z} = \sum_{i=1}^{n} a_i \mathbf{x}_i = \sum_{i=1}^{n} (\mathbf{y}_i, \mathbf{z})\mathbf{x}_i$$

2-4 DEFINITIONS OF TERMINOLOGIES IN MATRIX ANALYSIS

In this section, we review commonly encountered terminologies in matrix analysis. Specifically, we review definitions of eigenvalues, eigenvectors, norms, definiteness and semidefiniteness of quadratic forms and Hermitian forms, elementary transformations, the rank of a matrix, and similarity of matrices.

Eigenvalues of square matrix Let $\mathbf{A}$ be an $n \times n$ matrix. The determinant

$$|\mathbf{A} - \lambda\mathbf{I}| \qquad (\mathbf{I} = \text{identity matrix})$$

is called the *characteristic polynomial* of $\mathbf{A}$. It is an nth-degree polynomial in λ. The characteristic equation is given by $|\mathbf{A} - \lambda\mathbf{I}| = 0$, or

$$|\mathbf{A} - \lambda\mathbf{I}| = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix}$$

$$= (-)^n(\lambda^n + c_1\lambda^{n-1} + \cdots + c_{n-2}\lambda^2 + c_{n-1}\lambda + c_n) = 0$$

The n roots of the characteristic equation are the eigenvalues of $\mathbf{A}$. They are called the *characteristic roots*. Notice that a real $n \times n$ matrix $\mathbf{A}$ does not necessarily possess real eigenvalues. But since $|\mathbf{A} - \lambda\mathbf{I}| = 0$ is a polynomial with real coefficients, any complex eigenvalues must occur in conjugate pairs, namely, if $\alpha + j\beta$ is an eigenvalue, then $\alpha - j\beta$ is also an eigenvalue of $\mathbf{A}$. Notice also that for a diagonal matrix or a triangular matrix, n diagonal elements are the n eigenvalues of the matrix.

If we define the eigenvalues of $\mathbf{A}$ as $\lambda_i(\mathbf{A})$, and those of $\mathbf{A}^{-1}$ as $\mu_i(\mathbf{A}^{-1})$, then

$$\lambda_i^{-1}(\mathbf{A}) = \mu_i(\mathbf{A}^{-1}) \qquad (i = 1, 2, \ldots, n) \tag{2-6}$$

Namely, if λ_i is an eigenvalue of $\mathbf{A}$, then λ_i^{-1} is an eigenvalue of $\mathbf{A}^{-1}$. For example, if $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 2 & 2 \end{bmatrix}$$

then the eigenvalues of $\mathbf{A}$ are found from

$$\begin{vmatrix} 1 - \lambda & 3 \\ 2 & 2 - \lambda \end{vmatrix} = (\lambda + 1)(\lambda - 4) = 0$$

or

$$\lambda_1 = -1, \qquad \lambda_2 = 4$$

Hence,

$$\lambda_1^{-1}(\mathbf{A}) = \frac{1}{\lambda_1} = -1, \qquad \lambda_2^{-1}(\mathbf{A}) = \frac{1}{\lambda_2} = \frac{1}{4}$$

The inverse of $\mathbf{A}$ can be found as

$$\mathbf{A}^{-1} = \begin{bmatrix} -\frac{1}{2} & \frac{3}{4} \\ \frac{1}{2} & -\frac{1}{4} \end{bmatrix}$$

The eigenvalues of $\mathbf{A}^{-1}$ can be obtained from

$$\begin{vmatrix} -\frac{1}{2} - \mu & \frac{3}{4} \\ \frac{1}{2} & -\frac{1}{4} - \mu \end{vmatrix} = (\mu + 1)(\mu - \tfrac{1}{4}) = 0$$

or

$$\mu_1 = -1, \qquad \mu_2 = \tfrac{1}{4}$$

Hence,

$$\mu_1(\mathbf{A}^{-1}) = \mu_1 = -1, \qquad \mu_2(\mathbf{A}^{-1}) = \mu_2 = \tfrac{1}{4}$$

Thus we have demonstrated that Eq. (2-6) holds true.

Finally, it is noted that it is possible to prove that, for two square matrices $\mathbf{A}$ and $\mathbf{B}$

$$|\mathbf{AB} - \lambda\mathbf{I}| = |\mathbf{BA} - \lambda\mathbf{I}|$$

(For the proof see Prob. A-2-29.)

Eigenvectors of square matrix Any nonzero vector $\mathbf{x}_i$ such that $\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i$ is called an *eigenvector* associated with an eigenvalue λ_i of an $n \times n$ matrix $\mathbf{A}$. Since the components of $\mathbf{x}_i$ are determined from n linear homogeneous algebraic equations within a constant factor, if $\mathbf{x}_i$ is an eigenvector, then, for any scalar $\alpha \neq 0$, $\alpha\mathbf{x}_i$ is also an eigenvector. The eigenvector is called a *normalized* eigenvector if the length or absolute value of the eigenvector is unity.

Norms of matrix The concept of norms of a vector can be extended to matrices. There are several different definitions of norms of a matrix. Some of them follow:

(1) A norm $\|\mathbf{A}\|$ of an $n \times n$ matrix $\mathbf{A}$ may be defined by

$$\|\mathbf{A}\| = \min k$$

such that

$$\|\mathbf{Ax}\| \leqslant k\|\mathbf{x}\|$$

For the norm $[\mathbf{x}^*\mathbf{x}]^{1/2}$, this definition is equivalent to

$$\|\mathbf{A}\|^2 = \max_{\mathbf{x}}\{\mathbf{x}^*\mathbf{A}^*\mathbf{Ax};\ \mathbf{x}^*\mathbf{x} = 1\}$$

which means that $\|\mathbf{A}\|^2$ is the maximum of the "absolute value" of the vector $\mathbf{Ax}$ when $\mathbf{x}^*\mathbf{x} = 1$.

(2) A norm of an $n \times n$ matrix $\mathbf{A}$ may be defined by

$$\|\mathbf{A}\| = \sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|$$

where $|a_{ij}|$ is the absolute value of a_{ij}.

(3) A norm may be defined by

$$\|\mathbf{A}\| = \left(\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2\right)^{1/2}$$

(4) Another definition of a norm is given by

$$\|\mathbf{A}\| = \max_i\left\{\sum_{j=1}^{n}|a_{ij}|\right\}$$

Norms of an $n \times n$ matrix $\mathbf{A}$ have the following properties:

(1) $$\|\mathbf{A}\| = \|\mathbf{A}^*\|$$

(2) $$\|\mathbf{A}+\mathbf{B}\| \leqslant \|\mathbf{A}\| + \|\mathbf{B}\|$$

(3) $$\|\mathbf{AB}\| \leqslant \|\mathbf{A}\|\|\mathbf{B}\|$$

(4) $$\|\mathbf{Ax}\| \leqslant \|\mathbf{A}\|\|\mathbf{x}\|$$

Quadratic form, bilinear form, and Hermitian form For a real symmetric matrix $\mathbf{A}$ and a real n vector $\mathbf{x}$ the following form:

$$\mathbf{x}'\mathbf{Ax} = \sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}x_ix_j \qquad (a_{ji} = a_{ij})$$

is called a *real quadratic* form in x_i. $\mathbf{x}'\mathbf{Ax}$ is a real scalar quantity.

For an $n \times m$ real matrix $\mathbf{A}$, a real n vector $\mathbf{x}$, and a real m vector $\mathbf{y}$, the following form:

$$\mathbf{x}'\mathbf{Ay} = \sum_{i=1}^{n}\sum_{j=1}^{m}a_{ij}x_iy_j$$

is called a *real bilinear* form in x_i and y_j. $\mathbf{x}'\mathbf{Ay}$ is a real scalar quantity.

For a Hermitian matrix $\mathbf{A}$ and a complex n vector $\mathbf{x}$, the following form

$$\mathbf{x}^*\mathbf{Ax} = \sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}\bar{x}_ix_j \qquad (a_{ji} = \bar{a}_{ij})$$

is called a *Hermitian form*. Notice that the scalar quantity $\mathbf{x}^*\mathbf{Ax}$ is real, because

$$\overline{\mathbf{x}^*\mathbf{Ax}} = \mathbf{x}'\bar{\mathbf{A}}\bar{\mathbf{x}} = (\mathbf{x}'\bar{\mathbf{A}}\bar{\mathbf{x}})' = \bar{\mathbf{x}}'\bar{\mathbf{A}}'\mathbf{x} = \mathbf{x}^*\mathbf{Ax}$$

Among the three forms defined here, the quadratic form and Hermitian form are most important in the study of stability analysis of dynamic systems (see Chapter 8) and optimal control systems (see Chapter 9). In this book we shall therefore focus our attention mostly on the quadratic form and Hermitian form.

Definiteness and semidefiniteness A quadratic form $\mathbf{x}'\mathbf{Ax}$ where $\mathbf{A}$ is a real symmetric matrix (or a Hermitian form $\mathbf{x}^*\mathbf{Ax}$ where $\mathbf{A}$ is a Hermitian matrix) is said to be *positive definite* if

$$\mathbf{x}'\mathbf{Ax} > 0 \quad (\text{or } \mathbf{x}^*\mathbf{Ax} > 0) \quad \text{for } \mathbf{x} \neq \mathbf{0}$$
$$\mathbf{x}'\mathbf{Ax} = 0 \quad (\text{or } \mathbf{x}^*\mathbf{Ax} = 0) \quad \text{for } \mathbf{x} = \mathbf{0}$$

$\mathbf{x}'\mathbf{Ax}$ (or $\mathbf{x}^*\mathbf{Ax}$) is said to be positive semidefinite if

$$\mathbf{x}'\mathbf{Ax} \geqslant 0 \quad (\text{or } \mathbf{x}^*\mathbf{Ax} \geqslant 0) \quad \text{for } \mathbf{x} \neq \mathbf{0}$$
$$\mathbf{x}'\mathbf{Ax} = 0 \quad (\text{or } \mathbf{x}^*\mathbf{Ax} = 0) \quad \text{for } \mathbf{x} = \mathbf{0}$$

$\mathbf{x}'\mathbf{Ax}$ (or $\mathbf{x}^*\mathbf{Ax}$) is said to be *negative definite* if

$$\mathbf{x}'\mathbf{Ax} < 0 \quad (\text{or } \mathbf{x}^*\mathbf{Ax} < 0) \quad \text{for } \mathbf{x} \neq \mathbf{0}$$
$$\mathbf{x}'\mathbf{Ax} = 0 \quad (\text{or } \mathbf{x}^*\mathbf{Ax} = 0) \quad \text{for } \mathbf{x} = \mathbf{0}$$

$\mathbf{x}'\mathbf{Ax}$ (or $\mathbf{x}^*\mathbf{Ax}$) is said to be *negative semidefinite* if

$$\mathbf{x}'\mathbf{Ax} \leqslant 0 \quad (\text{or } \mathbf{x}^*\mathbf{Ax} \leqslant 0) \quad \text{for } \mathbf{x} \neq \mathbf{0}$$
$$\mathbf{x}'\mathbf{Ax} = 0 \quad (\text{or } \mathbf{x}^*\mathbf{Ax} = 0) \quad \text{for } \mathbf{x} = \mathbf{0}$$

If $\mathbf{x}'\mathbf{A}\mathbf{x}$ (or $\mathbf{x}^*\mathbf{A}\mathbf{x}$) can be of either sign, $\mathbf{x}'\mathbf{A}\mathbf{x}$ (or $\mathbf{x}^*\mathbf{A}\mathbf{x}$) is said to be *indefinite*.

Necessary and sufficient conditions for the quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ (or Hermitian form $\mathbf{x}^*\mathbf{A}\mathbf{x}$) to be positive definite, negative definite, positive semidefinite, or negative semidefinite were given by T. T. Sylvester. We shall postpone detailed discussions of Sylvester's criteria to Section 5–5.

Linear dependence and independence A set of k columns $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k$ of an $n \times m$ matrix $\mathbf{A}$ is linearly independent if and only if the equation

$$c_1\mathbf{A}_1 + c_2\mathbf{A}_2 + \cdots + c_k\mathbf{A}_k = \mathbf{0}$$

implies

$$c_1 = c_2 = \cdots = c_k = 0$$

If there exist elements c_i, not all zero, such that

$$c_1\mathbf{A}_1 + c_2\mathbf{A}_2 + \cdots + c_k\mathbf{A}_k = \mathbf{0}$$

then a set of k columns of a matrix $\mathbf{A}$ is linearly dependent. For example, the columns of the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & \sqrt{2} & 0 \\ 0 & 0 & 0 \\ 0 & 1 & \sqrt{3} \end{bmatrix}$$

are linearly dependent, since

$$(-\sqrt{2})\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + (1)\begin{bmatrix} \sqrt{2} \\ 0 \\ 1 \end{bmatrix} + \left(\frac{-1}{\sqrt{3}}\right)\begin{bmatrix} 0 \\ 0 \\ \sqrt{3} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

If k rows (or columns) of a matrix are linearly independent, then any subset of i rows ($i < k$) is linearly independent. If a set of k nonzero rows (or columns) of a matrix $\mathbf{A}$ is linearly dependent, then there exists a maximum subset of i rows (or columns) ($i < k$) which is linearly independent. The remaining $k - i$ rows (or columns) are linear combinations of i linearly independent rows (or columns). Consider the following matrix:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & -1 & 0 \\ 1 & 0 & 1 & 2 \\ 1 & 1 & 0 & 2 \end{bmatrix}$$

The four columns of $\mathbf{A}$ are not linearly independent, since the following equation:

$$c_1\begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} + c_2\begin{bmatrix} 2 \\ 1 \\ 0 \\ 1 \end{bmatrix} + c_3\begin{bmatrix} 3 \\ -1 \\ 1 \\ 0 \end{bmatrix} + c_4\begin{bmatrix} 4 \\ 0 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

can be satisfied by the nonzero c_i. To demonstrate this let us solve the following simultaneous equations

$$\begin{aligned} c_1 + 2c_2 + 3c_3 + 4c_4 &= 0 \\ c_2 - c_3 &= 0 \\ c_1 + c_3 + 2c_4 &= 0 \\ c_1 + c_2 + 2c_4 &= 0 \end{aligned}$$

which yield

$$c_1 = 1.5, \qquad c_2 = 0.5, \qquad c_3 = 0.5, \qquad c_4 = -1$$

We therefore see that the four columns of $\mathbf{A}$ are not linearly independent. However, the three columns of a submatrix of $\mathbf{A}$

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & -1 \\ 1 & 0 & 1 \end{bmatrix}$$

are linearly independent, since

$$c_1\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + c_2\begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} + c_3\begin{bmatrix} 3 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

implies that $c_1 = c_2 = c_3 = 0$. We can easily verify that any subset of these three columns is linearly independent.

Elementary transformations Suppose that an $n \times m$ matrix $\mathbf{A}$ is a linear transformation from a vector space V' to a vector space V. The effect of a change of basis in V is to replace $\mathbf{A}$ by $\mathbf{PA}$ where $\mathbf{P}$ is a nonsingular $n \times n$ matrix. The effect of a change of basis in V' is to replace $\mathbf{A}$ by $\mathbf{AQ}$, where $\mathbf{Q}$ is a nonsingular $m \times m$ matrix. By choosing suitable bases for V and V' the matrix $\mathbf{A}$ can be given the following simple form:

$$\mathbf{PAQ} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where $\mathbf{I}_r$ is the $r \times r$ identity matrix ($r \leqslant n$, $r \leqslant m$) and the remaining elements are zero. The matrix $\mathbf{A}$ can be given this simple form by performing a suitable sequence of elementary transformations of the following types:

(1) The interchange of any two basis vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ in V (or $\mathbf{y}_i$ and $\mathbf{y}_j$ in V').
(2) The replacement of a basis vector $\mathbf{x}_i$ by $c\mathbf{x}_i$ (or $\mathbf{y}_i$ by $c\mathbf{y}_i$) where c is a nonzero scalar.
(3) The replacement of a basis vector $\mathbf{x}_i$ by $\mathbf{x}_i + c\mathbf{x}_j$ (or $\mathbf{y}_i$ by $\mathbf{y}_i + c\mathbf{y}_j$) where c is a nonzero scalar.

The effect of elementary transformations is to carry out analogous operations called elementary row (or column) operations on the rows (or columns) of the matrix $\mathbf{A}$.

Elementary row or column operations The following row (or column) operations on a matrix $\mathbf{A}$ are called the *elementary row* (or column) operations.

(1) The interchange of any two rows (or columns).
(2) The multiplication of a row (or column) by a constant c where $c \neq 0$.
(3) The addition of one row (or column) multiplied by a constant to another row (or column).

An $n \times n$ matrix obtained from the $n \times n$ identity matrix $\mathbf{I}$ by means of one elementary row (or column) operation is called an *elementary* matrix. A few examples are

$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} c & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The elementary matrices are nonsingular. Notice that the inverse of an elementary matrix is also an elementary matrix. Clearly, elementary operations do not alter the number of linearly independent rows (or columns) of the matrix.

Any nonzero $n \times m$ matrix $\mathbf{A}$ can be reduced by elementary operations to a matrix of the form

$$\begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where $\mathbf{I}_r$ is the $r \times r$ identity matrix ($r \leqslant n$, $r \leqslant m$) and the remaining elements are zero. In the following, we shall prove the foregoing statement. Since $\mathbf{A}$ is a nonzero matrix, at least one element is nonzero. We can always bring a nonzero element in the (1, 1)th place by means of elementary operations. Also we can make this element equal to 1 by multiplying an appropriate constant. It is simple, by means of elementary operations, to transform all the elements in the first row and the first column except the (1, 1)th element equal to zero. Then matrix $\mathbf{A}$ becomes

$$\left[\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{A}_1 \end{array}\right]$$

If $\mathbf{A}_1 \neq \mathbf{0}$, we repeat the same process. Notice that this process will not affect the first row and the first column. If the same process is continued, then the transformed matrix finally becomes of the form

$$\begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

In the preceding process, we employed only elementary operations. Therefore, we may write

$$\mathbf{PAQ} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where both $\mathbf{P}$ and $\mathbf{Q}$ are products of elementary matrices.

We shall next show that every nonsingular matrix $\mathbf{A}$ is a product of elementary matrices. Since there exist matrices $\mathbf{P}$ and $\mathbf{Q}$ such that $\mathbf{PAQ} = \mathbf{I}$, where $\mathbf{P}$ and $\mathbf{Q}$ are products of elementary matrices, and therefore are nonsingular, we have $\mathbf{A} = \mathbf{P}^{-1}\mathbf{Q}^{-1}$. Since $\mathbf{P}^{-1}$ and $\mathbf{Q}^{-1}$ are products of inverses of elementary matrices and the inverse of an elementary matrix is also an elementary matrix, $\mathbf{P}^{-1}\mathbf{Q}^{-1} = \mathbf{A}$ is a product of elementary matrices. Hence we have proved that any nonsingular matrix is a product of elementary matrices.

Rank of matrix We shall first prove that the maximum number of linearly independent rows of a matrix $\mathbf{A}$ is equal to the maximum number of linearly independent columns. Let $\mathbf{A}$ be any nonzero $n \times m$ matrix and let $\mathbf{P}$ be an $n \times n$ nonsingular matrix and $\mathbf{Q}$ be an $m \times m$ nonsingular matrix such that

$$\mathbf{PAQ} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where $\mathbf{I}_r$ is the $r \times r$ identity matrix ($r \leqslant n$, $r \leqslant m$). The maximum number of linearly independent rows of $\mathbf{PAQ}$ is r, which is the dimension of the $r \times r$ identity matrix. The maximum number of linearly independent rows of $\mathbf{PA}$ is the same as that of $\mathbf{A}$, since $\mathbf{P}$ is a nonsingular matrix which is a product of elementary matrices. (Elementary operations do not alter the maximum number of linearly independent rows—or columns.) Since

$$\mathbf{PA} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{Q}^{-1} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where $\mathbf{Q}_{11}$ and $\mathbf{Q}_{12}$ are r-rowed matrices and since $\mathbf{Q}^{-1}$ is nonsingular, the m rows are linearly independent. Hence the maximum number of linearly independent rows of $\mathbf{PA}$ or equivalently $\mathbf{A}$, is r.

Consider next

$$\mathbf{Q}^*\mathbf{A}^*\mathbf{P}^* = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

The maximum number of linearly independent rows of $\mathbf{Q}^*\mathbf{A}^*$ is the same as that of $\mathbf{A}^*$. By a discussion similar to the preceding we find the maximum number of linearly independent rows of $\mathbf{Q}^*\mathbf{A}^*$ to be r. Hence the maximum number of linerly independent rows of $\mathbf{A}^*$ is r, which means that the maximum number of linearly independent columns of $\mathbf{A}$ is r. Thus we have proved that the maximum number of linearly independent rows of a matrix is equal to that of linearly independent columns.

We shall now define the rank of a matrix. A matrix $\mathbf{A}$ is called of rank m, if the maximum number of linearly independent rows (or columns) is m. Hence if there exists an $m \times m$ submatrix $\mathbf{M}$ of $\mathbf{A}$ such that $|\mathbf{M}| \neq 0$ and the determinant of every $r \times r$ submatrix (where $r \geqslant m + 1$) of $\mathbf{A}$ is zero, then the rank of $\mathbf{A}$ is m. (If the determinant of every $(m + 1) \times (m + 1)$ submatrix of $\mathbf{A}$ is zero, then any determinant of order s $(s > m + 1)$ is zero, since any determinant of order $s > m + 1$ can be expressed as a linear sum of determinants of order $m + 1$.) As an example, consider the following matrix:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & -1 & 0 \\ 1 & 0 & 1 & 2 \\ 1 & 1 & 0 & 2 \end{bmatrix}$$

A maximum submatrix whose determinant is not equal to zero is

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & -1 \\ 1 & 0 & 1 \end{bmatrix}$$

Hence the rank of $\mathbf{A}$ is three.

The computation of determinants for the determination of the rank of a matrix may be time-consuming. The computational effort can be reduced greatly by means of elementary row or column operations, since elementary operations do not alter the rank of a matrix. The invariance of the rank under elementary operations can be seen as follows: Since the rank of a matrix $\mathbf{A}$ is the maximum number of linearly independent rows (or columns), elementary operations of type (1) and type (2) clearly will not alter the rank of $\mathbf{A}$. Since the maximum number of linearly independent rows (or columns) is not altered by the addition of one row (or column) multiplied by a constant

to another row (or column), the elementary operation of type (3) also will not alter the rank of $\mathbf{A}$. Hence the rank of a matrix $\mathbf{A}$ is invariant under elementary operations. In order to determine the rank of a given matrix $\mathbf{A}$, we may therefore premultiply and/or postmultiply $\mathbf{A}$ by a series of elementary matrices and transform $\mathbf{A}$ into a simpler form with a number of zero elements so that the rank can be found immediately.

Rank of product of matrices Consider an $n \times m$ matrix $\mathbf{A}$ and an $m \times p$ matrix $\mathbf{B}$. Assume that the rank of $\mathbf{A}$ is r. The rank of the product $\mathbf{AB}$ cannot exceed the rank of $\mathbf{A}$ or rank of $\mathbf{B}$, namely,

$$\text{rank } \mathbf{AB} \leqslant \text{rank } \mathbf{A}$$

$$\text{rank } \mathbf{AB} \leqslant \text{rank } \mathbf{B}$$

To prove this notice that nonsingular matrices $\mathbf{P}$ and $\mathbf{Q}$ exist such that

$$\mathbf{PAQ} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

or

$$\mathbf{PA} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{Q}^{-1} = \begin{bmatrix} \mathbf{F} \\ \text{--} \\ \mathbf{0} \end{bmatrix}$$

where $\mathbf{F}$ is a matrix of rank r. Consider next

$$\mathbf{PAB} = \begin{bmatrix} \mathbf{F} \\ \text{---} \\ \mathbf{0} \end{bmatrix} \mathbf{B} = \begin{bmatrix} \mathbf{H} \\ \text{--} \\ \mathbf{0} \end{bmatrix}$$

Since $\mathbf{F}$ has r nonzero rows, $\mathbf{H}$ has at most r nonzero rows. Hence

$$\text{rank } \mathbf{PAB} \leqslant r$$

Since $\mathbf{P}$ is a nonsingular matrix

$$\text{rank } \mathbf{PAB} = \text{rank } \mathbf{AB}$$

Therefore

$$\text{rank } \mathbf{AB} \leqslant \text{rank } \mathbf{A}$$

Similarly, we obtain

$$\text{rank } \mathbf{B}^*\mathbf{A}^* \leqslant \text{rank } \mathbf{B}^*$$

Since rank $\mathbf{B}^*\mathbf{A}^*$ = rank $\mathbf{AB}$ and rank $\mathbf{B}^*$ = rank $\mathbf{B}$, we obtain

$$\text{rank } \mathbf{AB} \leqslant \text{rank } \mathbf{B}$$

Equivalence of two matrices An $n \times m$ matrix $\mathbf{A}$ is said to be *equivalent* to an $n \times m$ matrix $\mathbf{B}$ if and only if there exist an $n \times n$ nonsingular matrix $\mathbf{P}$ and an $m \times m$ nonsingular matrix $\mathbf{Q}$ such that $\mathbf{A} = \mathbf{PBQ}$. This is equivalent to saying that $\mathbf{A}$ and $\mathbf{B}$ are equivalent if and only if they have the

same rank. To prove this statement notice that premultiplication and postmultiplication by nonsingular matrices, which are products of elementary matrices, do not alter the rank. Therefore, if $\mathbf{A} = \mathbf{PBQ}$, then the rank of $\mathbf{A}$ is the same as that of $\mathbf{B}$. On the other hand if $\mathbf{A}$ and $\mathbf{B}$ have the same rank, say r, then there exist an $n \times n$ nonsingular matrix $\mathbf{G}$ and an $m \times m$ nonsingular matrix $\mathbf{H}$ such that

$$\mathbf{A} = \mathbf{G}\begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\mathbf{H}$$

and also there exist an $n \times n$ nonsingular matrix $\mathbf{S}$ and an $m \times m$ nonsingular matrix $\mathbf{T}$ such that

$$\mathbf{B} = \mathbf{S}\begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\mathbf{T}$$

or

$$\begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \mathbf{S}^{-1}\mathbf{B}\mathbf{T}^{-1}$$

Hence,

$$\mathbf{A} = \mathbf{G}\mathbf{S}^{-1}\mathbf{B}\mathbf{T}^{-1}\mathbf{H} = (\mathbf{G}\mathbf{S}^{-1})\,\mathbf{B}(\mathbf{T}^{-1}\mathbf{H})$$

Since $\mathbf{GS}^{-1}$ and $\mathbf{T}^{-1}\mathbf{H}$ are nonsingular matrices, we may put $\mathbf{GS}^{-1} = \mathbf{P}$ and $\mathbf{T}^{-1}\mathbf{H} = \mathbf{Q}$. Thus, if $\mathbf{A}$ and $\mathbf{B}$ have the same rank, then $\mathbf{A} = \mathbf{PBQ}$, or $\mathbf{A}$ and $\mathbf{B}$ are equivalent.

Note that every matrix $\mathbf{A}$ is equivalent to a unique $n \times m$ matrix of the form

$$\mathbf{J}_{rnm} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

In other words, the matrices $\mathbf{J}_{rnm}$ for various r, n, m are a set of canonical forms under equivalence.

Similar matrices Let us next consider a matrix representing a linear transformation f of the vector space V into itself. Let us define a new basis $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_n$ in V by the equations

$$\begin{aligned}
\hat{\mathbf{x}}_1 &= q_{11}\mathbf{x}_1 + q_{12}\mathbf{x}_2 + \cdots + q_{1n}\mathbf{x}_n \\
\hat{\mathbf{x}}_2 &= q_{21}\mathbf{x}_1 + q_{22}\mathbf{x}_2 + \cdots + q_{2n}\mathbf{x}_n \\
&\cdots \qquad\quad \cdots \qquad\quad \cdots \\
\hat{\mathbf{x}}_n &= q_{n1}\mathbf{x}_1 + q_{n2}\mathbf{x}_2 + \cdots + q_{nn}\mathbf{x}_n
\end{aligned}$$

The matrix $\mathbf{Q} = (q_{ij})$ is a nonsingular matrix. Suppose we define

$$\mathbf{Q}^{-1} = \mathbf{P} = (p_{ij})$$

then

$$\begin{aligned}
\mathbf{x}_1 &= p_{11}\hat{\mathbf{x}}_1 + p_{12}\hat{\mathbf{x}}_2 + \cdots + p_{1n}\hat{\mathbf{x}}_n \\
\mathbf{x}_2 &= p_{21}\hat{\mathbf{x}}_1 + p_{22}\hat{\mathbf{x}}_2 + \cdots + p_{2n}\hat{\mathbf{x}}_n \\
&\cdots \qquad\quad \cdots \qquad\quad \cdots \\
\mathbf{x}_n &= p_{n1}\hat{\mathbf{x}}_1 + p_{n2}\hat{\mathbf{x}}_2 + \cdots + p_{nn}\hat{\mathbf{x}}_n
\end{aligned}$$

Let f have a matrix $\mathbf{A}$ relative to the basis $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Then f has the matrix $\mathbf{P'AQ'}$ relative to the basis $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_n$.

Two $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ are called *similar*, if a nonsingular matrix $\mathbf{P}$ exists such that

$$\mathbf{P}^{-1}\mathbf{AP} = \mathbf{B}$$

The matrix $\mathbf{B}$ is said to be obtained from $\mathbf{A}$ by a *similarity transformation*, in which $\mathbf{P}$ is the transformation matrix. Notice that $\mathbf{A}$ can be obtained from $\mathbf{B}$ by a similarity transformation with a transformation matrix $\mathbf{P}^{-1}$, since

$$\mathbf{A} = \mathbf{PBP}^{-1} = (\mathbf{P}^{-1})^{-1}\,\mathbf{B}(\mathbf{P}^{-1})$$

A change of basis in V, therefore, replaces $\mathbf{A}$ by a similar matrix $\mathbf{B}$.

If $\mathbf{A}$ can be reduced to a similar matrix which has a simple form, then important properties of $\mathbf{A}$ can be readily observed. A property of a matrix is called *invariant* if it is possessed by all similar matrices. For example, the determinant and the characteristic polynomial are invariant under a similarity transformation as shown:

$$\begin{aligned}
|\mathbf{B}| &= |\mathbf{P}^{-1}\mathbf{AP}| = |\mathbf{P}^{-1}||\mathbf{A}||\mathbf{P}| \\
&= |\mathbf{A}||\mathbf{P}^{-1}||\mathbf{P}| = |\mathbf{A}||\mathbf{P}^{-1}\mathbf{P}| = |\mathbf{A}||\mathbf{I}| = |\mathbf{A}|
\end{aligned}$$

and

$$\begin{aligned}
|\mathbf{B} - \lambda\mathbf{I}| &= |\mathbf{P}^{-1}\mathbf{AP} - \lambda\mathbf{I}| \\
&= |\mathbf{P}^{-1}\mathbf{AP} - \mathbf{P}^{-1}(\lambda\mathbf{I})\mathbf{P}| \\
&= |\mathbf{P}^{-1}(\mathbf{A} - \lambda\mathbf{I})\mathbf{P}| \\
&= |\mathbf{P}^{-1}||\mathbf{A} - \lambda\mathbf{I}||\mathbf{P}| \\
&= |\mathbf{A} - \lambda\mathbf{I}||\mathbf{P}^{-1}||\mathbf{P}| \\
&= |\mathbf{A} - \lambda\mathbf{I}|
\end{aligned}$$

Notice that the trace of a matrix is also invariant under similarity transformation, or

$$\operatorname{tr}\mathbf{A} = \operatorname{tr}(\mathbf{P}^{-1}\mathbf{AP})$$

But the property of symmetry of a matrix is not invariant.

Notice that only invariant properties of matrices present intrinsic characteristics of the class of similar matrices. To determine the invariant properties of a matrix $\mathbf{A}$ we examine the Jordan canonical form of $\mathbf{A}$. (For the Jordan canonical form, refer to Section 3–1. Here we shall state only

that similarity of two matrices can be defined in terms of the Jordan canonical form; and the necessary and sufficient condition that $n \times n$ matrices A and B be similar is that the Jordan canonical form of **A** and that of **B** be identical.)

EXERCISE PROBLEMS AND SOLUTIONS

Problem A-2-1 Suppose that **A** and **B** are $n \times n$ matrices. Show that the square matrix **AB** can be resolved into a sum of a symmetric matrix and a skew symmetric matrix.

Solution We can write

$$2\mathbf{AB} = (\mathbf{AB} + \mathbf{B'A'}) + (\mathbf{AB} - \mathbf{B'A'})$$

where

$\mathbf{AB} + \mathbf{B'A'}$ is symmetric

$\mathbf{AB} - \mathbf{B'A'}$ is skew symmetric

Problem A-2-2 Prove that a square matrix **A** possesses the inverse if and only if $|\mathbf{A}| \neq 0$.

Solution If $|\mathbf{A}| \neq 0$, then $\text{adj}\,\mathbf{A}/|\mathbf{A}|$ is the inverse of **A**, since

$$\frac{\text{adj}\,\mathbf{A}}{|\mathbf{A}|}\mathbf{A} = \frac{|\mathbf{A}|\mathbf{I}}{|\mathbf{A}|} = \mathbf{I}$$

also

$$\mathbf{A}\frac{\text{adj}\,\mathbf{A}}{|\mathbf{A}|} = \frac{|\mathbf{A}|\mathbf{I}}{|\mathbf{A}|} = \mathbf{I}$$

On the other hand, if **B** is the inverse of **A**, then $\mathbf{AB} = \mathbf{I}$ and we obtain

$$|\mathbf{A}||\mathbf{B}| = |\mathbf{I}| = 1$$

Hence

$$|\mathbf{A}| \neq 0$$

Problem A-2-3 Prove that

$$(\text{adj}\,\mathbf{A})(\text{adj}\,\mathbf{B}) = \text{adj}\,\mathbf{BA}$$

where A and B are nonsingular.

Solution

$$\begin{aligned}
&(\text{adj}\,\mathbf{A})(\text{adj}\,\mathbf{B})\,\mathbf{BA} \\
&= (\text{adj}\,\mathbf{A})(\det\mathbf{B})\,\mathbf{IA} \\
&= (\det\mathbf{B})(\text{adj}\,\mathbf{A})\,\mathbf{A} \\
&= (\det\mathbf{B})(\det\mathbf{A})\,\mathbf{I} \\
&= (\det\mathbf{BA})\,\mathbf{I} \\
&= (\text{adj}\,\mathbf{BA})\,\mathbf{BA}
\end{aligned}$$

Hence

$$(\text{adj}\,\mathbf{A})(\text{adj}\,\mathbf{B}) = \text{adj}\,\mathbf{BA}$$

Problem A-2-4 Vandermonde's determinant of order n has the following factorization

$$\begin{vmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_1^{n-1} & x_2^{n-1} & \cdots & x_n^{n-1} \end{vmatrix} = \prod_{1 \leqslant i < j \leqslant n} (x_j - x_i)$$

Verify this when the x_i are distinct.

Solution For $n = 2$ the given relationship is obviously true. To prove this relationship for $n \geqslant 3$, let us use induction. Assume that the given relationship is true for $n - 1$ and deduce that it is true for n.

Expanding the determinant about the last column, we obtain a polynomial in x_n of degree $n - 1$. Let us define

$$\begin{vmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_1^{n-1} & x_2^{n-1} & \cdots & x_n^{n-1} \end{vmatrix} = f(x_n)$$

The coefficient of the highest-degree term in x_n is

$$\begin{vmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_{n-1} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_1^{n-2} & x_2^{n-2} & \cdots & x_{n-1}^{n-2} \end{vmatrix}$$

which is, by assumption, equal to

$$\prod_{1 \leqslant i < j \leqslant n-1} (x_j - x_i)$$

Since

$$f(x_1) = f(x_2) = \cdots = f(x_{n-1}) = 0$$

we can express $f(x_n)$ in the following form:

$$f(x_n) = \Big[\prod_{1 \leqslant i < j \leqslant n-1} (x_j - x_i)\Big](x_n - x_1)(x_n - x_2)\cdots(x_n - x_{n-1})$$

it follows that

$$f(x_n) = \prod_{1 \leqslant i < j \leqslant n} (x_j - x_i)$$

This completes the proof.

Problem A-2-5 Obtain the inverse of the following matrix **A**:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & a_1 \\ 0 & 1 & 0 & 0 & a_2 \\ 0 & 0 & 1 & 0 & a_3 \\ 0 & 0 & 0 & 1 & a_4 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Solution

$$\mathbf{A}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & -a_1 \\ 0 & 1 & 0 & 0 & -a_2 \\ 0 & 0 & 1 & 0 & -a_3 \\ 0 & 0 & 0 & 1 & -a_4 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Problem A-2-6 Obtain $\mathbf{A}^{-1}$ where **A** is given by

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -\alpha^3 & \alpha^3 & -\beta^3 & \beta^3 \\ \alpha & -\alpha & \beta & -\beta \\ -\alpha^2 & -\alpha^2 & -\beta^2 & -\beta^2 \end{bmatrix}$$

Solution

$$\mathbf{A}^{-1} = \frac{1}{2(\beta^2 - \alpha^2)} \begin{bmatrix} \beta^2 & \alpha^{-1} & \beta^2\alpha^{-1} & 1 \\ \beta^2 & -\alpha^{-1} & -\beta^2\alpha^{-1} & 1 \\ -\alpha^2 & -\beta^{-1} & -\alpha^2\beta^{-1} & -1 \\ -\alpha^2 & \beta^{-1} & \alpha^2\beta^{-1} & -1 \end{bmatrix}$$

Problem A-2-7 Find the inverse of the following matrix

$$\mathbf{A} = \begin{bmatrix} \lambda + \frac{s^3}{3} & \frac{s^2}{2} & -\frac{s^3}{3} & -\frac{s^2}{2} \\ \frac{s^2}{2} & \lambda + s & -\frac{s^2}{2} & -s \\ 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & \lambda \end{bmatrix}$$

Solution

$$\mathbf{A}^{-1} = \frac{1}{\Delta} \begin{bmatrix} \lambda^2(\lambda + s) & -\frac{\lambda^2}{2}s^2 & \lambda\left(\frac{\lambda}{3}s^3 + \frac{s^4}{12}\right) & \frac{\lambda^2}{2}s^2 \\ -\frac{\lambda^2}{2}s^2 & \lambda^2\left(\lambda + \frac{s^3}{3}\right) & \frac{\lambda^2}{2}s^2 & \frac{\lambda}{12}(12\lambda s + s^4) \\ 0 & 0 & \frac{\lambda}{12}(12\lambda^2 + 12\lambda s + 4\lambda s^3 + s^4) & 0 \\ 0 & 0 & 0 & \frac{\lambda}{12}(12\lambda^2 + 12\lambda s + 4\lambda s^3 + s^4) \end{bmatrix}$$

where

$$\Delta = \lambda^4 + \lambda^3 s + \lambda^3\frac{s^3}{3} + \lambda^2\frac{s^4}{12}$$

Problem A-2-8 Consider the following n simultaneous linear equations:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \cdots \quad & \quad \cdots \quad \cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned}$$

which are equivalent to

$$\mathbf{Ax} = \mathbf{B}$$

where

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$

Assuming that $|\mathbf{A}| \neq 0$, obtain the solution of the simultaneous equations in the form of Cramer's rule.

Solution **x** is obtained as

$$\begin{aligned} \mathbf{x} &= \mathbf{A}^{-1}\mathbf{B} \\ &= \frac{1}{|\mathbf{A}|} \begin{bmatrix} \Delta_{11} & \Delta_{21} & \cdots & \Delta_{n1} \\ \Delta_{12} & \Delta_{22} & \cdots & \Delta_{n2} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \Delta_{1n} & \Delta_{2n} & \cdots & \Delta_{nn} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{bmatrix} \\ &= \frac{1}{|\mathbf{A}|} \begin{bmatrix} \sum_{i=1}^{n} b_i \Delta_{i1} \\ \sum_{i=1}^{n} b_i \Delta_{i2} \\ \cdot \\ \cdot \\ \cdot \\ \sum_{i=1}^{n} b_i \Delta_{in} \end{bmatrix} \end{aligned}$$

Hence

$$x_j = \frac{1}{|\mathbf{A}|}\sum_{i=1}^{n} b_i \Delta_{ij} = \frac{1}{|\mathbf{A}|} \begin{vmatrix} a_{11} & \cdots & b_1 & \cdots & a_{1n} \\ a_{21} & \cdots & b_2 & \cdots & a_{2n} \\ \cdot & & \cdot & & \cdot \\ \cdot & & \cdot & & \cdot \\ a_{n1} & \cdots & b_n & \cdots & a_{nn} \end{vmatrix} = \frac{|\mathbf{A}^{(j)}|}{|\mathbf{A}|}$$

where $|\mathbf{A}^{(j)}|$ is the determinant obtained from $\mathbf{A}$ by replacing the jth column by the column matrix $\mathbf{B}$. This gives the solution for x_j $(j = 1, 2, \ldots, n)$ and is Cramer's rule.

Problem A-2-9 Prove that for $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$

$$\begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{vmatrix} = |\mathbf{A} + \mathbf{B}| \cdot |\mathbf{A} - \mathbf{B}|$$

Solution

$$\begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{vmatrix} = \begin{vmatrix} \mathbf{A}+\mathbf{B} & \mathbf{B}+\mathbf{A} \\ \mathbf{B} & \mathbf{A} \end{vmatrix} = \begin{vmatrix} \mathbf{A}+\mathbf{B} & \mathbf{0} \\ \mathbf{B} & \mathbf{A}-\mathbf{B} \end{vmatrix} = |\mathbf{A} + \mathbf{B}| \cdot |\mathbf{A} - \mathbf{B}|$$

Note that for $n \times n$ matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ the determinant

$$\begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{vmatrix}$$

is, in general, neither equal to $|\mathbf{AD} - \mathbf{BC}|$ nor equal to $|\mathbf{A}||\mathbf{D}| - |\mathbf{B}||\mathbf{C}|$.

Problem A-2-10 Obtain $\frac{d}{dt}|\mathbf{A}(t)|$ where

$$|\mathbf{A}(t)| = \begin{vmatrix} a_{11}(t) & a_{12}(t) & \cdots & a_{1n}(t) \\ a_{21}(t) & a_{22}(t) & \cdots & a_{2n}(t) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n1}(t) & a_{n2}(t) & \cdots & a_{nn}(t) \end{vmatrix}$$

Solution The derivative of the determinant $|\mathbf{A}(t)|$ is given by a linear combination of all minors of $|\mathbf{A}(t)|$ as follows:

$$\frac{d}{dt}|\mathbf{A}(t)| = \sum_{i=1}^{n} \begin{vmatrix} a_{11}(t) & a_{12}(t) & \cdots & a_{1n}(t) \\ \vdots & \vdots & & \vdots \\ \frac{d}{dt}a_{i1}(t) & \frac{d}{dt}a_{i2}(t) & \cdots & \frac{d}{dt}a_{in}(t) \\ \vdots & \vdots & & \vdots \\ a_{n1}(t) & a_{n2}(t) & \cdots & a_{nn}(t) \end{vmatrix}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_{ij} \frac{d}{dt} a_{ij}(t)$$

Problem A-2-11 Suppose that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are a set of orthonormal vectors, or

$$\begin{aligned} (\mathbf{x}_i, \mathbf{x}_j) &= 1 \quad (i = j) \\ &= 0 \quad (i \neq j) \\ (i, j &= 1, 2, \ldots, n) \end{aligned}$$

then for vectors $\mathbf{y}$ and $\mathbf{z}$ defined by

$$\mathbf{y} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_n\mathbf{x}_n$$
$$\mathbf{z} = b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \cdots + b_n\mathbf{x}_n$$

show that

$$(\mathbf{y}, \mathbf{z}) = \bar{a}_1 b_1 + \bar{a}_2 b_2 + \cdots + \bar{a}_n b_n$$

and

$$(\mathbf{y}, \mathbf{x}_i) = \bar{a}_i \qquad (i = 1, 2, \ldots, n)$$

Solution Since

$$(\mathbf{y}, \mathbf{z}) = \left(\sum_{i=1}^{n} a_i\mathbf{x}_i, \sum_{j=1}^{n} b_j\mathbf{x}_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \bar{a}_i b_j(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=1}^{n} \bar{a}_i b_i$$

we obtain

$$(\mathbf{y}, \mathbf{z}) = \bar{a}_1 b_1 + \bar{a}_2 b_2 + \cdots + \bar{a}_n b_n$$

By substituting $\mathbf{z} = \mathbf{x}_i$ we obtain

$$(\mathbf{y}, \mathbf{x}_i) = \bar{a}_i \qquad (i = 1, 2, \ldots, n)$$

Problem A-2-12 For $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$

$$\text{tr}\,(\mathbf{AB}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ji} = \text{tr}\,(\mathbf{BA})$$

Show that

$$\text{tr}\,(\mathbf{AB}') = \text{tr}\,(\mathbf{A}'\mathbf{B}) = \text{tr}\,(\mathbf{BA}')$$

Solution Since $\text{tr}\,\mathbf{A} = \text{tr}\,\mathbf{A}'$, we have

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ij} = \text{tr}\,(\mathbf{AB}') = \text{tr}\,(\mathbf{BA}')$$

also

$$\text{tr}\,(\mathbf{BA}') = \text{tr}\,(\mathbf{A}'\mathbf{B})$$

Problem A-2-13 Show that if a scalar quantity f is defined by

$$f = \mathbf{x}'\mathbf{A}\mathbf{x}$$

then

$$\mathbf{x}'(\mathbf{A} - \mathbf{A}')\mathbf{x} = 0$$

Solution Since f is a scalar quantity, the transpose of f is equal to f,

or

$$\mathbf{x}'\mathbf{A}\mathbf{x} = f = f' = \mathbf{x}'\mathbf{A}'\mathbf{x}$$

Hence,

$$\mathbf{x}'(\mathbf{A} - \mathbf{A}')\mathbf{x} = 0$$

Problem A-2-14 Given the characteristic equation

$$f(\lambda) = |\mathbf{A} - \lambda\mathbf{I}| = (-)^n(\lambda^n + c_1\lambda^{n-1} + \cdots + c_{n-1}\lambda + c_n) = 0$$

show that

$$\lambda_1\lambda_2 \cdots \lambda_n = (-)^n c_n = |\mathbf{A}|$$

and

$$\lambda_1 + \lambda_2 + \cdots + \lambda_n = \text{tr } \mathbf{A} = (-)^{n-1} c_1$$

Solution The characteristic equation has roots $\lambda_1, \lambda_2, \ldots, \lambda_n$. In a factored form the characteristic equation becomes

$$(\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n) = \prod_{i=1}^{n} (\lambda - \lambda_i) = 0$$

The only term which does not contain λ is the term consisting of the product of $-\lambda_i$ $(i = 1, 2, \ldots, n)$, and it is equal to c_n. Therefore,

$$c_n = (-)^n \prod_{i=1}^{n} (\lambda_i)$$

or

$$\lambda_1 \lambda_2 \cdots \lambda_n = (-)^n c_n$$

To show that $(-)^n c_n = |\mathbf{A}|$, write $|\mathbf{A} - \lambda\mathbf{I}|$ in the following form:

$$|\mathbf{A} - \lambda\mathbf{I}| = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix}$$

$$= \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} + (\text{polynomial in } \lambda \text{ of degree } n-1)\lambda$$

Note that in the expansion of $|\mathbf{A} - \lambda\mathbf{I}|$ the only term that does not involve λ is $|\mathbf{A}|$. Hence we obtain

$$|\mathbf{A}| = (-1)^n c_n$$

Notice that

$$-(\lambda_1 + \lambda_2 + \cdots + \lambda_n) = (-)^n c_1$$

Notice also that the trace of $\mathbf{A}$ is invariant under similarity transformation. Therefore,

$$\text{tr } \mathbf{A} = \text{tr}(\mathbf{P}^{-1}\mathbf{A}\mathbf{P}) = \lambda_1 + \lambda_2 + \cdots + \lambda_n$$

Hence

$$\lambda_1 + \lambda_2 + \cdots + \lambda_n = \text{tr } \mathbf{A} = (-)^{n-1} c_1$$

Problem A-2-15 Given the following $n \times n$ matrix

$$\mathbf{A} = \begin{bmatrix} -a_1 & 1 & 0 & \cdots & 0 \\ -a_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -a_{n-1} & 0 & 0 & \cdots & 1 \\ -a_n & 0 & 0 & \cdots & 0 \end{bmatrix}$$

show that the characteristic polynomial is

$$|\mathbf{A} - \lambda\mathbf{I}| = (-)^n(\lambda^n + a_1\lambda^{n-1} + a_2\lambda^{n-2} + \cdots + a_{n-1}\lambda + a_n)$$

Solution

$$|\mathbf{A} - \lambda\mathbf{I}| = \begin{vmatrix} -a_1 - \lambda & 1 & 0 & \cdots & 0 \\ -a_2 & -\lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -a_{n-1} & 0 & 0 & \cdots & 1 \\ -a_n & 0 & 0 & \cdots & -\lambda \end{vmatrix}$$

$$= (-a_1 - \lambda) \begin{vmatrix} -\lambda & 1 & 0 & \cdots & 0 \\ 0 & -\lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & -\lambda \end{vmatrix}$$

$$+ (-) \begin{vmatrix} -a_2 & 1 & 0 & \cdots & 0 \\ -a_3 & -\lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -a_{n-1} & 0 & 0 & \cdots & 1 \\ -a_n & 0 & 0 & \cdots & -\lambda \end{vmatrix}$$

$$= (-a_1 - \lambda)(-\lambda)^{n-1} + (-)^2 a_2 \begin{vmatrix} -\lambda & 1 & 0 & \cdots & 0 \\ 0 & -\lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda \end{vmatrix}$$

$$+ (-)^2 \begin{vmatrix} -a_3 & 1 & 0 & \cdots & 0 \\ -a_4 & -\lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -a_n & 0 & 0 & \cdots & -\lambda \end{vmatrix}$$

$$= (-a_1 - \lambda)(-\lambda)^{n-1} + (-)^n a_2\lambda^{n-2} + (-)^n a_3\lambda^{n-3} + \cdots + (-)^n a_{n-1}\lambda + (-)^n a_n$$

$$= (-)^n(\lambda^n + a_1\lambda^{n-1} + a_2\lambda^{n-2} + a_3\lambda^{n-3} + \cdots + a_{n-1}\lambda + a_n)$$

Problem A-2-16 Obtain eigenvectors of the following matrix:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

Solution The characteristic equation is

$$|\mathbf{A}-\lambda\mathbf{I}|=\begin{vmatrix}-\lambda & 1 & 0 & 0\\ 0 & -\lambda & 1 & 0\\ 0 & 0 & -\lambda & 1\\ 1 & 0 & 0 & -\lambda\end{vmatrix}=\lambda^4-1=0$$

Hence $\lambda_1=1$, $\lambda_2=-1$, $\lambda_3=j$, $\lambda_4=-j$. Let the eigenvector for λ_i be

$$\mathbf{x}_i=\begin{bmatrix}x_{1i}\\ x_{2i}\\ x_{3i}\\ x_{4i}\end{bmatrix}$$

Then

$$\begin{bmatrix}0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 1 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}x_{1i}\\ x_{2i}\\ x_{3i}\\ x_{4i}\end{bmatrix}=\lambda_i\begin{bmatrix}x_{1i}\\ x_{2i}\\ x_{3i}\\ x_{4i}\end{bmatrix}$$

it follows that

$$\begin{bmatrix}x_{2i}\\ x_{3i}\\ x_{4i}\\ x_{1i}\end{bmatrix}=\begin{bmatrix}\lambda_i x_{1i}\\ \lambda_i x_{2i}\\ \lambda_i x_{3i}\\ \lambda_i x_{4i}\end{bmatrix}$$

Hence the eigenvector for λ_i is given by

$$\begin{bmatrix}x_{1i}\\ x_{2i}\\ x_{3i}\\ x_{4i}\end{bmatrix}=c_i\begin{bmatrix}1\\ \lambda_i\\ \lambda_i^2\\ \lambda_i^3\end{bmatrix}$$

where c_i is an arbitrary constant different from zero.

Let $\mathbf{T}$ be a matrix consisting of four independent eigenvectors of $\mathbf{A}$. Then, an example of $\mathbf{T}$ is given by

$$\mathbf{T}=\begin{bmatrix}1 & 1 & 1 & 1\\ \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4\\ \lambda_1^2 & \lambda_2^2 & \lambda_3^2 & \lambda_4^2\\ \lambda_1^3 & \lambda_2^3 & \lambda_3^3 & \lambda_4^3\end{bmatrix}=\begin{bmatrix}1 & 1 & 1 & 1\\ 1 & -1 & j & -j\\ 1 & 1 & -1 & -1\\ 1 & -1 & -j & j\end{bmatrix}$$

Problem A-2-17 Show that for the following matrix $\mathbf{A}$:

$$\mathbf{A}=\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}$$

there does not exist any nonsingular matrix $\mathbf{T}$ such that

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T}=\mathbf{I}$$

Solution The eigenvalues of $\mathbf{A}$ are obtained from

$$\begin{vmatrix}1-\lambda & 1\\ 0 & 1-\lambda\end{vmatrix}=(1-\lambda)^2=0$$

Thus the two eigenvalues are equal. Suppose that there exists a nonsingular matrix $\mathbf{T}$ such that

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T}=\mathbf{I}$$

Premultiplying with $\mathbf{T}$

$$\mathbf{A}\mathbf{T}=\mathbf{T}$$

or

$$\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}\begin{bmatrix}t_{11} & t_{12}\\ t_{21} & t_{22}\end{bmatrix}=\begin{bmatrix}t_{11} & t_{12}\\ t_{21} & t_{22}\end{bmatrix}$$

Solving for t_{ij}, we obtain the following:

$$\mathbf{T}=\begin{bmatrix}t_{11} & t_{12}\\ 0 & 0\end{bmatrix}$$

Thus, $\mathbf{T}$ is singular. Since $\mathbf{T}$ is assumed to be nonsingular, this is a contradiction. Therefore, we conclude that there does not exist any nonsingular matrix $\mathbf{T}$ such that

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T}=\mathbf{I}$$

Problem A-2-18 Prove that every eigenvalue λ_i $(i=1,2,\ldots,n)$ of a unitary matrix (or orthogonal matrix) $\mathbf{A}$ has the magnitude of unity. (Hence the real eigenvalues of a unitary matrix must be either 1 or -1.)

Solution Since orthogonal matrices are unitary, we shall prove this for a unitary matrix. Notice that there exists a column vector $\mathbf{x}$ such that

$$\mathbf{A}\mathbf{x}=\lambda_i\mathbf{x}\qquad(i=1,2,\ldots,n)$$

If we take the conjugate transpose of both sides of this equation, then

$$\mathbf{x}^*\mathbf{A}^*=\bar{\lambda}_i\mathbf{x}^*$$

Therefore,

$$\mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x}=\bar{\lambda}_i\mathbf{x}^*\lambda_i\mathbf{x}$$

Since $\mathbf{A}$ is a unitary matrix, we have $\mathbf{A}^*\mathbf{A}=\mathbf{I}$. Hence

$$\mathbf{x}^*\mathbf{x}=\lambda_i\bar{\lambda}_i\mathbf{x}^*\mathbf{x}$$

or

$$(1-\lambda_i\bar{\lambda}_i)\mathbf{x}^*\mathbf{x}=0$$

Since $\mathbf{x}^*\mathbf{x}\neq 0$, we obtain

$$\lambda_i\bar{\lambda}_i=1$$

or

$$|\lambda_i|^2 = 1$$

it follows that the magnitude of every eigenvalue of a unitary matrix is unity.

Problem A-2-19 Prove that for a normal matrix $\mathbf{A}$

$$\mathbf{Ax} = \lambda\mathbf{x}$$

implies that

$$\mathbf{A}^*\mathbf{x} = \bar{\lambda}\mathbf{x}$$

Solution

$$\begin{aligned}\|(\mathbf{A} - \lambda\mathbf{I})\mathbf{x}\|^2 &= ((\mathbf{A} - \lambda\mathbf{I})\mathbf{x}, (\mathbf{A} - \lambda\mathbf{I})\mathbf{x}) \\ &= (\mathbf{x}, (\mathbf{A} - \lambda\mathbf{I})^*(\mathbf{A} - \lambda\mathbf{I})\mathbf{x}) \\ &= (\mathbf{x}, (\mathbf{A} - \lambda\mathbf{I})(\mathbf{A} - \lambda\mathbf{I})^*\mathbf{x}) \\ &= ((\mathbf{A} - \lambda\mathbf{I})^*\mathbf{x}, (\mathbf{A} - \lambda\mathbf{I})^*\mathbf{x}) \\ &= \|(\mathbf{A} - \lambda\mathbf{I})^*\mathbf{x}\|^2\end{aligned}$$

From this we see that

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$$

implies that

$$(\mathbf{A} - \lambda\mathbf{I})^*\mathbf{x} = \mathbf{0}$$

or

$$\mathbf{A}^*\mathbf{x} = \bar{\lambda}\mathbf{x}$$

Problem A-2-20 Prove that the eigenvalues of an $n \times n$ Hermitian matrix and of an $n \times n$ real symmetric matrix are real. Prove also that the eigenvalues of a skew Hermitian matrix and of a real skew symmetric matrix are either zero or purely imaginary.

Solution Let us define any eigenvalue of an $n \times n$ Hermitian matrix $\mathbf{A}$ by $\lambda = \alpha + j\beta$. There exists a vector $\mathbf{x} \neq \mathbf{0}$ such that

$$\mathbf{Ax} = (\alpha + j\beta)\mathbf{x}$$

Then

$$\mathbf{x}^*\mathbf{A}^* = (\alpha - j\beta)\mathbf{x}^*$$

Since $\mathbf{A}$ is Hermitian, $\mathbf{A}^* = \mathbf{A}$. Therefore,

$$\mathbf{x}^*\mathbf{Ax} = (\alpha - j\beta)\mathbf{x}^*\mathbf{x}$$

On the other hand, since $\mathbf{Ax} = (\alpha + j\beta)\mathbf{x}$, we have

$$\mathbf{x}^*\mathbf{Ax} = (\alpha + j\beta)\mathbf{x}^*\mathbf{x}$$

Hence we obtain

$$[(\alpha - j\beta) - (\alpha + j\beta)]\mathbf{x}^*\mathbf{x} = 0$$

or

$$-2j\beta\mathbf{x}^*\mathbf{x} = 0$$

Since $\mathbf{x}^*\mathbf{x} \neq 0$ (for $\mathbf{x} \neq \mathbf{0}$), we conclude that

$$\beta = 0$$

This proves that any eigenvalue of an $n \times n$ Hermitian matrix $\mathbf{A}$ is real. It follows that the eigenvalues of a real symmetric matrix are real, since it is Hermitian.

To prove the second half of the problem notice that if $\mathbf{B}$ is skew Hermitian, then $j\mathbf{B}$ is Hermitian. Hence the eigenvalues of $j\mathbf{B}$ are real, which implies that the eigenvalues of $\mathbf{B}$ are either zero or purely imaginary.

The eigenvalues of a real skew symmetric matrix are also either zero or purely imaginary, since a real skew symmetric matrix is skew Hermitian. (Notice that purely imaginary eigenvalues always occur in conjugate pairs in the real skew symmetric matrix, since the coefficients of the characteristic equation are real. Notice also that an $n \times n$ real skew symmetric matrix is singular if n is odd, since such a matrix must include at least one zero eigenvalue.)

Problem A-2-21 Prove that eigenvectors associated with distinct eigenvalues of an $n \times n$ Hermitian matrix $\mathbf{A}$ are orthogonal.

Solution Let us define n eigenvectors of $\mathbf{A}$ by the $\mathbf{x}_i$. Then,

$$\mathbf{Ax}_i = \lambda_i\mathbf{x}_i \qquad (i = 1, 2, \ldots, n)$$

where the λ_i are the eigenvalues of $\mathbf{A}$. (The λ_i are real.) Since

$$\begin{aligned}\lambda_i(\mathbf{x}_i, \mathbf{x}_j) &= (\lambda_i\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{Ax}_i, \mathbf{x}_j) = (\mathbf{x}_i, \mathbf{A}^*\mathbf{x}_j) = (\mathbf{x}_i, \mathbf{Ax}_j) \\ &= (\mathbf{x}_i, \lambda_j\mathbf{x}_j) = \lambda_j(\mathbf{x}_i, \mathbf{x}_j)\end{aligned}$$

we obtain

$$(\lambda_i - \lambda_j)(\mathbf{x}_i, \mathbf{x}_j) = 0$$

Since $\lambda_i \neq \lambda_j$, we have

$$(\mathbf{x}_i, \mathbf{x}_j) = 0$$

This proves that eigenvectors associated with distinct eigenvalues of an $n \times n$ Hermitian matrix $\mathbf{A}$ are orthogonal.

Problem A-2-22 Show that any real quadratic form can be represented by $\mathbf{x}'\mathbf{Ax}$ or $(\mathbf{x}, \mathbf{Ax})$ or $(\mathbf{Ax}, \mathbf{x})$, where $\mathbf{A}$ is a real symmetric matrix. (The matrix $\mathbf{A}$ is called the *coefficient* matrix of the quadratic form.)

Solution An arbitrary real quadratic form can be represented by

$$\sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{i<j}^{n} a_{ij}x_ix_j$$

Let us put

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{12} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Then we obtain

$$\sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{i<j}^{n} a_{ij}x_ix_j = \mathbf{x}'\mathbf{Ax} = (\mathbf{x}, \mathbf{Ax}) = (\mathbf{Ax}, \mathbf{x})$$

By definition, $\mathbf{A}$ is symmetric. In addition, the a_{ij} are real. Hence, $\mathbf{A}$ is a real symmetric matrix. (Notice that real quadratic forms can be expressed in terms of real symmetric matrices and real vectors, whereas Hermitian forms can be expressed in terms of Hermitian matrices and complex vectors.)

Problem A-2-23 Show that if $\mathbf{A}$ is nonsingular, then $\mathbf{A}^*\mathbf{A}$ is positive definite.

Solution

$$\mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x} = \|\mathbf{A}\mathbf{x}\|^2 \geqslant 0$$

Since the matrix $\mathbf{A}$ is nonsingular, $\mathbf{A}\mathbf{x} = \mathbf{0}$ means that $\mathbf{x} = \mathbf{0}$. Therefore $\mathbf{A}^*\mathbf{A}$ is positive definite.

Problem A-2-24 Show that if a matrix $\mathbf{A}$ is positive definite, then $\mathbf{A}^{-1}$ is also positive definite.

Solution If $\mathbf{A}$ is positive definite, then we can write $\mathbf{A} = \mathbf{B}^*\mathbf{B}$ for some nonsingular matrix $\mathbf{B}$. Then $\mathbf{A}^{-1}$ can be written as $\mathbf{A}^{-1} = \mathbf{B}^{-1}(\mathbf{B}^{-1})^*$. This implies that $\mathbf{A}^{-1}$ is also positive definite.

Problem A-2-25 Determine the rank of the following matrices;

(1) $$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & -1 \\ 0 & 1 & -5 & 4 \\ 1 & 3 & -2 & 3 \\ 2 & 5 & 1 & 2 \end{bmatrix}$$

(2) $$\mathbf{A} = \begin{bmatrix} 1 & -1 & 4 & 4 \\ 2 & 2 & 0 & 4 \\ 4 & 3 & 2 & 9 \end{bmatrix}$$

Solution

(1) Rank is 2.
(2) Rank is 2.

Problem A-2-26 Show that if a square matrix $\mathbf{A}$ is reduced to the identity matrix by a sequence of transformations, then the same sequence of operations performed on the identity matrix produces $\mathbf{A}^{-1}$.

Solution Let $\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_n$ be transformations. If n transformations reduce $\mathbf{A}$ to $\mathbf{I}$, then

$$(\mathbf{T}_n \cdots \mathbf{T}_2\mathbf{T}_1)\mathbf{A} = \mathbf{I}$$

Postmultiplying $\mathbf{A}^{-1}$ into both sides of the preceding equation, we obtain

$$(\mathbf{T}_n \cdots \mathbf{T}_2\mathbf{T}_1)(\mathbf{A}\mathbf{A}^{-1}) = \mathbf{I}\mathbf{A}^{-1}$$

or

$$(\mathbf{T}_n \cdots \mathbf{T}_2\mathbf{T}_1)\mathbf{I} = \mathbf{A}^{-1}$$

Problem A-2-27 Show that the following matrix $\mathbf{A}$ can be reduced to the identity matrix by a sequence of elementary operations.

$$\mathbf{A} = \begin{bmatrix} 1 & -2 \\ 1 & 1 \end{bmatrix}$$

Show also that the same sequence of elementary operations performed on the identity matrix produces $\mathbf{A}^{-1}$.

Solution A sequence of transformed matrices is shown in the left column. The same sequence of elementary operations performed on $\mathbf{I}$ is shown in the right column.

$$\mathbf{A} = \begin{bmatrix} 1 & -2 \\ 1 & 1 \end{bmatrix} \qquad \mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 0 & -3 \\ 1 & 1 \end{bmatrix} \qquad \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \qquad \begin{bmatrix} -\frac{1}{3} & \frac{1}{3} \\ 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad \begin{bmatrix} -\frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{I} \qquad \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \\ -\frac{1}{3} & \frac{1}{3} \end{bmatrix} = \mathbf{A}^{-1}$$

Problem A-2-28 Given matrices $\mathbf{A}$, $\mathbf{P}$, and $\mathbf{Q}$

where

$\mathbf{A} = n \times m$ matrix

$\mathbf{P} = n \times n$ nonsingular matrix

$\mathbf{Q} = m \times m$ nonsingular matrix

prove that

$$\text{rank } \mathbf{PAQ} = \text{rank } \mathbf{PA} = \text{rank } \mathbf{AQ} = \text{rank } \mathbf{A}$$

Solution Notice that

$$\text{rank } \mathbf{PA} \leqslant \text{rank } \mathbf{A} = \text{rank } \mathbf{P}^{-1}(\mathbf{PA}) \leqslant \text{rank } \mathbf{PA}$$

This implies that rank $\mathbf{PA}$ = rank $\mathbf{A}$. Similarly, rank $\mathbf{AQ}$ = rank $\mathbf{A}$. Hence,

$$\text{rank } \mathbf{PAQ} = \text{rank } \mathbf{P}(\mathbf{AQ}) = \text{rank } \mathbf{AQ} = \text{rank } \mathbf{A}$$

Problem A-2-29 Given two $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$, prove that the eigenvalues of $\mathbf{AB}$ and those of $\mathbf{BA}$ are identical.

Solution We shall consider first the case where $\mathbf{A}$ (or $\mathbf{B}$) is nonsingular. In this case clearly

$$|\mathbf{BA} - \lambda\mathbf{I}| = |\mathbf{A}^{-1}(\mathbf{AB})\mathbf{A} - \lambda\mathbf{I}| = |\mathbf{A}^{-1}(\mathbf{AB} - \lambda\mathbf{I})\mathbf{A}|$$
$$= |\mathbf{A}^{-1}||\mathbf{AB} - \lambda\mathbf{I}||\mathbf{A}| = |\mathbf{AB} - \lambda\mathbf{I}|$$

We shall next consider the case where both $\mathbf{A}$ and $\mathbf{B}$ are singular. There exist $n \times n$ nonsingular matrices $\mathbf{P}$ and $\mathbf{Q}$ such that

$$\mathbf{PAQ} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where $\mathbf{I}_r$ is the $r \times r$ identity matrix and r is the rank of $\mathbf{A}$, $r < n$. We have

$$|\mathbf{BA} - \lambda\mathbf{I}| = |\mathbf{Q}^{-1}\mathbf{BAQ} - \lambda\mathbf{I}| = |\mathbf{Q}^{-1}\mathbf{BP}^{-1}\mathbf{PAQ} - \lambda\mathbf{I}|$$
$$= \left| \begin{bmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} - \lambda\mathbf{I} \right|$$

where

$$\mathbf{Q}^{-1}\mathbf{B}\mathbf{P}^{-1} = \begin{bmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{bmatrix}$$

Then

$$\begin{aligned} |\mathbf{BA} - \lambda\mathbf{I}| &= \left| \begin{bmatrix} \mathbf{G}_{11} & \mathbf{0} \\ \mathbf{G}_{21} & \mathbf{0} \end{bmatrix} - \lambda\mathbf{I} \right| \\ &= \begin{vmatrix} \mathbf{G}_{11} - \lambda\mathbf{I}_r & \mathbf{0} \\ \mathbf{G}_{21} & -\lambda\mathbf{I}_{n-r} \end{vmatrix} \\ &= |\mathbf{G}_{11} - \lambda\mathbf{I}_r| \cdot |-\lambda\mathbf{I}_{n-r}| \end{aligned}$$

Also

$$\begin{aligned} |\mathbf{AB} - \lambda\mathbf{I}| &= |\mathbf{PABP}^{-1} - \lambda\mathbf{I}| = |\mathbf{PAQQ}^{-1}\mathbf{BP}^{-1} - \lambda\mathbf{I}| \\ &= \left| \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{bmatrix} - \lambda\mathbf{I} \right| \\ &= \left| \begin{bmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} - \lambda\mathbf{I} \right| \\ &= \begin{vmatrix} \mathbf{G}_{11} - \lambda\mathbf{I}_r & \mathbf{G}_{12} \\ \mathbf{0} & -\lambda\mathbf{I}_{n-r} \end{vmatrix} \\ &= |\mathbf{G}_{11} - \lambda\mathbf{I}_r| \cdot |-\lambda\mathbf{I}_{n-r}| \end{aligned}$$

Hence we have proved that

$$|\mathbf{BA} - \lambda\mathbf{I}| = |\mathbf{AB} - \lambda\mathbf{I}|$$

PROBLEMS

Problem B-2-1 Show that if $\mathbf{A}$ is a nonsingular triangular matrix, then $\mathbf{A}^{-1}$ is also a triangular matrix.

Problem B-2-2 Show that the following 4×4 determinant:

$$|\mathbf{A}| = \begin{vmatrix} a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \\ d_1 & d_2 & d_3 & d_4 \end{vmatrix}$$

can be expanded into the following form: (This is called *Laplace's expansion by the minors*.)

$$\begin{aligned} |\mathbf{A}| = &\begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix} \begin{vmatrix} c_3 & c_4 \\ d_3 & d_4 \end{vmatrix} - \begin{vmatrix} a_1 & a_2 \\ c_1 & c_2 \end{vmatrix} \begin{vmatrix} b_3 & b_4 \\ d_3 & d_4 \end{vmatrix} \\ &+ \begin{vmatrix} a_1 & a_2 \\ d_1 & d_2 \end{vmatrix} \begin{vmatrix} b_3 & b_4 \\ c_3 & c_4 \end{vmatrix} + \begin{vmatrix} b_1 & b_2 \\ c_1 & c_2 \end{vmatrix} \begin{vmatrix} a_3 & a_4 \\ d_3 & d_4 \end{vmatrix} \\ &- \begin{vmatrix} b_1 & b_2 \\ d_1 & d_2 \end{vmatrix} \begin{vmatrix} a_3 & a_4 \\ c_3 & c_4 \end{vmatrix} + \begin{vmatrix} c_1 & c_2 \\ d_1 & d_2 \end{vmatrix} \begin{vmatrix} a_3 & a_4 \\ b_3 & b_4 \end{vmatrix} \end{aligned}$$

Problem B-2-3 Show that if $\mathbf{A} = \mathbf{A}'$, then

$$\mathbf{x}'\mathbf{B}'\mathbf{A}\mathbf{x} + \mathbf{x}'\mathbf{A}\mathbf{B}\mathbf{x} = 2\mathbf{x}'(\mathbf{A}\mathbf{B})'\mathbf{x}$$

Note that

$$(\mathbf{x}'\mathbf{A}\mathbf{B}\mathbf{x})' = \mathbf{x}'\mathbf{A}\mathbf{B}\mathbf{x}$$

Problem B-2-4 Verify that

$$\begin{bmatrix} 0 & 0 & 0 & a_4 \\ 1 & 0 & 0 & a_3 \\ 0 & 1 & 0 & a_2 \\ 0 & 0 & 1 & a_1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_1 & 1 & 0 & 0 \\ a_2 & 0 & 1 & 0 \\ a_3 & 0 & 0 & 1 \\ a_4 & 0 & 0 & 0 \end{bmatrix}$$

and then show that

$$\mathbf{S}^{-1} \begin{bmatrix} 0 & 0 & \cdots & 0 & a_n \\ 1 & 0 & \cdots & 0 & a_{n-1} \\ 0 & 1 & \cdots & 0 & a_{n-2} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & a_2 \\ 0 & 0 & \cdots & 1 & a_1 \end{bmatrix} \mathbf{S} = \begin{bmatrix} a_1 & 1 & 0 & \cdots & 0 & 0 \\ a_2 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ a_{n-1} & 0 & 0 & \cdots & 0 & 1 \\ a_n & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

where

$$\mathbf{S} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

Problem B-2-5 Prove that for an $n \times n$ real matrix $\mathbf{A}$ and a real n vector $\mathbf{x}$

$$(\mathbf{Ax}, \mathbf{Ax}) = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} a_{ij}x_i \right)^2$$

Prove also that for real n vectors $\mathbf{x}$ and $\mathbf{y}$

$$\begin{aligned} (\mathbf{x}, \mathbf{y})^2 &= \sum_{i=1}^{n} x_i^2 y_i^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} x_i x_j y_i y_j \\ &= \sum_{i=1}^{n} x_i^2 y_i^2 + 2 \sum_{i=1}^{n-1} \sum_{k=1}^{n-i} x_i x_{i+k} y_i y_{i+k} \end{aligned}$$

Problem B-2-6 Prove that a necessary and sufficient condition that a real matrix $\mathbf{A}$ be orthogonal is that

$$(\mathbf{Ax}, \mathbf{Ay}) = (\mathbf{x}, \mathbf{y})$$

for all real vectors $\mathbf{x}$ and $\mathbf{y}$.

Problem B-2-7 Suppose that $\mathbf{A}$ is a transformation from an n-dimensional vector space V to an n-dimensional vector space V'. Explain that if $|\mathbf{A}| = 1$, then $\mathbf{A}$ is a rotation or a product of rotations. If $|\mathbf{A}| = -1$, $\mathbf{A}$ is a reflec-

tion or a product of rotations and one reflection (or odd numbered reflections) and, therefore, the orientation is reversed.

Problem B-2-8 Prove that

$$\|\mathbf{x}+\mathbf{y}\|^2+\|\mathbf{x}-\mathbf{y}\|^2=2(\|\mathbf{x}\|^2+\|\mathbf{y}\|^2)$$

Problem B-2-9 Suppose that vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ are linearly independent and vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ are also linearly independent. Prove that if $\mathbf{y}_i$ $(1 \leqslant i \leqslant n)$ can be expressed as a linear combination of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ and if $\mathbf{x}_j$ $(1 \leqslant j \leqslant m)$ can be expressed as a linear combination of $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$, then m must be equal to n.

Problem B-2-10 Prove that if λ is an eigenvalue of an orthogonal matrix $\mathbf{A}$, then so is λ^{-1}. Similarly, prove that if λ is an eigenvalue of an unitary matrix $\mathbf{B}$, then so is $\bar{\lambda}^{-1}$.

Problem B-2-11 Prove that the characteristic equation of

$$\mathbf{A}=\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a_{22} & a_{23} & \cdots & a_{2n} \\ 0 & 0 & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{bmatrix}$$

is given by

$$|\mathbf{A}-\lambda\mathbf{I}|=(a_{11}-\lambda)(a_{22}-\lambda)\cdots(a_{nn}-\lambda)=0$$

(Hence the characteristic equation of a triangular matrix is determined by its main diagonal elements.)

Problem B-2-12 Assume that $\mathbf{A}$ is an $n \times n$ matrix possessing n distinct eigenvalues. Show that the matrix

$$\mathbf{A}-\lambda_k\mathbf{I}$$

is of rank $n-1$, where λ_k is the kth eigenvalue of $\mathbf{A}$.

Problem B-2-13 Prove that for $n \times m$ matrices $\mathbf{A}$ and $\mathbf{B}$

$$\text{rank}\,(\mathbf{A}+\mathbf{B}) \leqslant \text{rank}\,\mathbf{A}+\text{rank}\,\mathbf{B}$$

Problem B-2-14 Prove that for an $n \times m$ matrix $\mathbf{A}$ and $m \times r$ matrix $\mathbf{B}$

$$\text{rank}\,\mathbf{AB}+m \geqslant \text{rank}\,\mathbf{A}+\text{rank}\,\mathbf{B}$$

REFERENCES

2–1 Mirsky, L., *An Introduction to Linear Algebra*, Oxford: Clarendon Press, 1955.

2–2 Bellman, R., *Introduction to Matrix Analysis*, New York: McGraw-Hill Book Company, Inc., 1960.

2–3 Turnbull, H. W., and A. C. Aitken, *An Introduction to the Theory of Canonical Matrices*, London: Blackie and Son Ltd., 1932.

2–4 Halmos, P. R., *Finite Dimensional Vector Spaces*, Princeton, N. J.: D. Van Nostrand Company, Inc., 1958.

2–5 Kalman, R. E., and J. E. Bertram, "Control System Analysis and Design Via the Second Method of Lyapunov: I. Continuous Time Systems," *ASME J. Basic Engineering* (1960), 371–93.

3

Coordinate Transformation

3-1 INTRODUCTION

For a given dynamic system we may define a state vector in a number of different ways. A set of the components of a state vector, or a set of state variables, is the smallest collection of numbers which must be specified at time $t = t_0$ in order to be able to predict uniquely the behavior of the system for any time $t \geqslant t_0$ for any input belonging to the given input set in which all elements are known for $t \geqslant t_0$. Hence the dimension of the state vector is the same, however it is defined. The different ways of defining a state vector, therefore, can be viewed as transformation of the coordinates of the system.

In order to gain insight into the dynamic behavior of a system, it is helpful to make a coordinate transformation so that the coefficient matrix of the state vector will be in a canonical form exhibiting directly its eigenvalues on the main diagonal. Such a canonical form is called the *Jordan* canonical form.

Jordan canonical form An $n \times n$ matrix $\mathbf{A}$ is said to be in the *Jordan* canonical form if

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & & & 0 \\ & \mathbf{A}_2 & & \\ & & \ddots & \\ 0 & & & \mathbf{A}_s \end{bmatrix}$$

where the $\mathbf{A}_i$ are $m_i \times m_i$ matrices of the form

$$\mathbf{A}_i = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{bmatrix}$$

(λ in $\mathbf{A}_i$ and that in $\mathbf{A}_j$ may or may not be the same) and $m_1 + m_2 + \cdots + m_s = n$. In the matrix $\mathbf{A}$ each eigenvalue λ appears on the diagonal a number of times equal to its multiplicity. For example, in a 7×7 matrix $\mathbf{A}$, if $m_1 = 3$, $m_2 = 2$, $m_3 = 1$, $m_4 = 1$ and the eigenvalues of $\mathbf{A}$ are $\lambda_1, \lambda_1, \lambda_1, \lambda_1, \lambda_1, \lambda_6, \lambda_7$, then the Jordan canonical form is

$$\left[\begin{array}{ccc:cc:c:c} \lambda_1 & 1 & 0 & & & & 0 \\ 0 & \lambda_1 & 1 & & & & \\ 0 & 0 & \lambda_1 & & & & \\ \hdashline & & & \lambda_1 & 1 & & \\ & & & 0 & \lambda_1 & & \\ \hdashline & & & & & \lambda_6 & \\ \hdashline 0 & & & & & & \lambda_7 \end{array}\right]$$

Notice that a diagonal matrix is a special case of the Jordan canonical form.

Consider a linear dynamic system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$$

where $\mathbf{x}$ is a state vector (n vector) and $\mathbf{A}$ is an $n \times n$ constant matrix. Any state vector $\mathbf{y}$ defined for this system is related to $\mathbf{x}$ by

$$\mathbf{x} = \mathbf{P}\mathbf{y}$$

where $\mathbf{P} = n \times n$ nonsingular matrix

Among a number of state vectors which can be defined for this system a convenient one for investigating the system's dynamic behavior is a particular state vector $\mathbf{y}$ which yields the system differential equation in the following form:

$$\dot{\mathbf{y}} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}\mathbf{y} = \mathbf{J}\mathbf{y}$$

where $\mathbf{J} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ is in the simplest possible form, such as the diagonal form or more generally the Jordan canonical form. It is always possible to find a nonsingular matrix $\mathbf{P}$ that will transform an arbitrary square matrix $\mathbf{A}$ into the Jordan canonical form through the similarity transformation $\mathbf{J} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$.

The matrix **J** possesses the same invariant characteristics as **A** but is much easier to manipulate. Hence transforming **A** into the simplest possible form makes it possible to investigate intrinsic properties of the dynamic system with the minimum of manipulating labor.

In this chapter we discuss in detail transformations of the coefficient matrix **A** into the Jordan canonical form, which includes the diagonal form as a special case. In Section 3–2 we present preliminary discussions on diagonalization of matrices. Specifically, we shall diagonalize the coefficient matrix of the vector matrix differential equation with distinct eigenvalues into the diagonal form. In Section 3–3 we obtain the necessary and sufficient condition that a matrix be similar to a diagonal matrix. Also several theorems related to the diagonalization of matrices are presented. In Section 3–4 we treat systems with multiple eigenvectors. We shall obtain transformation matrices such that in the transformed equation the coefficient matrix is in the Jordan canonical form. Section 3–5 presents a brief discussion of machine computation of transformation matrices. Section 3–6 gives a summary of transformation matrices that reduce the coefficient matrix of the second order system into the Jordan canonical form. In Section 3–7 we treat systems described by nth-order scalar differential equations. We shall first transform them into state space equations and then obtain transformation matrices which transform the coefficient matrix into the Jordan canonical form. Finally, in Section 3–8 we discuss motions of linear free dynamic systems.

3-2 PRELIMINARY STUDY—SYSTEMS WITH DISTINCT EIGENVALUES

Consider the system described by

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}, \qquad \mathbf{x}(0) = \mathbf{c}$$

If we can find a suitable nonsingular matrix **P** such that the transformation $\mathbf{x} = \mathbf{P}\mathbf{y}$ transforms the given system into

$$\dot{\mathbf{y}} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}\mathbf{y}, \qquad \mathbf{y}(0) = \mathbf{P}^{-1}\mathbf{c} = \boldsymbol{\xi}$$

where

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

then the nth-order vector matrix differential equation (or the corresponding n simultaneous first-order differential equations) can be transformed into n first-order differential equations of the form

$$\dot{y}_i = \lambda_i y_i, \qquad y_i(0) = \xi_i \qquad (i = 1, 2, \ldots, n)$$

which can be solved independently of each other. Such a process is called *diagonalization of the coefficient matrix* **A**.

In Section 3–3, we shall prove that for an $n \times n$ matrix **A** with n distinct eigenvectors such diagonalization is possible if we take the n eigenvectors of **A** as coordinate axes and transform the vector **x** into vector **y**. (If **A** involves multiple eigenvectors, then **A** does not possess n linearly independent eigenvectors, and **A** cannot be similar to a diagonal matrix.)

In this section we shall present more or less preliminary discussions of diagonalization of matrices by use of a system with n distinct eigenvalues. (Such a system has n distinct eigenvectors.)

Consider the following set of n first-order differential equations:

$$\begin{aligned}
\dot{x}_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\
\dot{x}_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\
&\cdots \quad \cdots \quad \cdots \\
\dot{x}_n &= a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n \\
x_1(0) &= c_1, \qquad x_2(0) = c_2, \quad \ldots, \quad x_n(0) = c_n
\end{aligned} \tag{3–1}$$

where the a_{ij} are constant. Equations (3–1) can be written in the form of a vector matrix differential equation as follows:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}, \qquad \mathbf{x}(0) = \mathbf{c} \tag{3–2}$$

where

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \cdot \\ \cdot \\ \cdot \\ x_n(t) \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

Here we assume that **A** is an $n \times n$ constant matrix possessing n distinct eigenvalues, $\lambda_1, \lambda_2, \ldots, \lambda_n$.

Assume that the solution of Eq. (3–1) or Eq. (3–2) is given by

$$\begin{aligned}
x_1(t) &= c_1 e^{\lambda t} \\
x_2(t) &= c_2 e^{\lambda t} \\
&\cdots \\
x_n(t) &= c_n e^{\lambda t}
\end{aligned} \tag{3–3}$$

or

$$\mathbf{x}(t) = \mathbf{c}e^{\lambda t} \tag{3–4}$$

Substitution of Eq. (3–3) into Eq. (3–1) yields

$$\begin{aligned}
(a_{11} - \lambda)c_1 + a_{12}c_2 + \cdots + a_{1n}c_n &= 0 \\
a_{21}c_1 + (a_{22} - \lambda)c_2 + \cdots + a_{2n}c_n &= 0 \\
&\cdots \quad \cdots \quad \cdots \\
a_{n1}c_1 + a_{n2}c_2 + \cdots + (a_{nn} - \lambda)c_n &= 0
\end{aligned} \tag{3–5}$$

Equations (3–5) have a nontrivial solution if and only if

$$\begin{vmatrix} a_{11}-\lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22}-\lambda & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn}-\lambda \end{vmatrix} = 0 \tag{3–6}$$

or

$$|\mathbf{A} - \lambda\mathbf{I}| = 0$$

Equation (3–6) is the characteristic equation, or eigenvalue equation, its roots being characteristic roots or eigenvalues.

Eigenvectors We shall show that if a new coordinate system is chosen, each coordinate axis being an eigenvector, the variables in the original n simultaneous equations can be separated.

If ith eigenvalue λ_i is substituted into Eq. (3–5), then

$$\begin{aligned} (a_{11}-\lambda_i)c_1 + a_{12}c_2 + \cdots + a_{1n}c_n &= 0 \\ a_{21}c_1 + (a_{22}-\lambda_i)c_2 + \cdots + a_{2n}c_n &= 0 \\ \cdots \quad \cdots \quad \cdots & \\ a_{n1}c_1 + a_{n2}c_2 + \cdots + (a_{nn}-\lambda_i)c_n &= 0 \end{aligned} \tag{3–7}$$

The eigenvalue λ_i satisfies Eq. (3–6). For such a case, we cannot determine constants like $c_1, c_2, \ldots, c_n$ but only ratios like c_j/c_1. Referring to the determinant in the following equation,

$$\begin{vmatrix} a_{11}-\lambda_i & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22}-\lambda_i & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn}-\lambda_i \end{vmatrix} = 0 \tag{3–8}$$

if we define the cofactors of the first row as $A_{i11}, A_{i12}, \ldots, A_{i1n}$ and if $\lambda_i = \lambda_1$ is substituted into Eq. (3–8) and the determinant is expanded in terms of the first row, then we obtain

$$a_{11}A_{111} + a_{12}A_{112} + \cdots + a_{1n}A_{11n} = \lambda_1 A_{111}$$

If the first row of the determinant in Eq. (3–8) is replaced by the mth row where $m = 2, 3, \ldots, n$ (the value of the determinant is still zero) and the resulting determinant is then expanded in terms of the first row, we obtain

$$a_{m1}A_{111} + a_{m2}A_{112} + \cdots + a_{mn}A_{11n} = \lambda_1 A_{11m}$$
$$(m = 2, 3, \ldots, n)$$

Thus, for λ_1 we can get altogether n such relationships. Since the same applies to any λ_i, we can get a total of n^2 equations of the following form:

$$a_{j1}A_{i11} + a_{j2}A_{i12} + \cdots + a_{jn}A_{i1n} = \lambda_i A_{i1j} \tag{3–9}$$
$$(i, j = 1, 2, 3, \ldots, n)$$

We shall use Eq. (3–9) later. From this analysis we see that $c_1, c_2, \ldots, c_n$ can be obtained as follows:

$$\begin{aligned} c_1 &= KA_{i11} \\ c_2 &= KA_{i12} \\ &\cdots \\ c_n &= KA_{i1n} \end{aligned} \tag{3–10}$$

where K is a constant. Hence, the x_k defined in Eq. (3–3) become

$$\begin{aligned} x_1(t) &= KA_{i11}e^{\lambda_i t} \\ x_2(t) &= KA_{i12}e^{\lambda_i t} \\ &\cdots \\ x_n(t) &= KA_{i1n}e^{\lambda_i t} \end{aligned} \tag{3–11}$$

Equations (3–11) hold for each eigenvalue $\lambda_1, \lambda_2, \ldots, \lambda_n$. The general solution of Eq. (3–1) is then given by

$$\begin{aligned} x_1(t) &= K_1A_{111}e^{\lambda_1 t} + K_2A_{211}e^{\lambda_2 t} + \cdots + K_nA_{n11}e^{\lambda_n t} \\ x_2(t) &= K_1A_{112}e^{\lambda_1 t} + K_2A_{212}e^{\lambda_2 t} + \cdots + K_nA_{n12}e^{\lambda_n t} \\ \cdots \quad & \quad \cdots \quad \quad \cdots \\ x_n(t) &= K_1A_{11n}e^{\lambda_1 t} + K_2A_{21n}e^{\lambda_2 t} + \cdots + K_nA_{n1n}e^{\lambda_n t} \end{aligned} \tag{3–12}$$

where the K_i are constants to be determined from the initial conditions. If the initial conditions are given as

$$\begin{aligned} x_1(0) &= K_1A_{111} \\ x_2(0) &= K_1A_{112} \\ &\cdots \\ x_n(0) &= K_1A_{11n} \end{aligned} \tag{3–13}$$

substitution of Eq. (3–13) into the following, which is obtained by letting $t = 0$ in Eq. (3–12):

$$\begin{aligned} x_1(0) &= K_1A_{111} + K_2A_{211} + \cdots + K_nA_{n11} \\ x_2(0) &= K_1A_{112} + K_2A_{212} + \cdots + K_nA_{n12} \\ \cdots \quad & \quad \cdots \quad \quad \cdots \\ x_n(0) &= K_1A_{11n} + K_2A_{21n} + \cdots + K_nA_{n1n} \end{aligned}$$

results in

$$K_2 = K_3 = \cdots = K_n = 0$$

This means that under such an initial condition the solution is given by

$$\begin{aligned} x_1(t) &= K_1 A_{111} e^{\lambda_1 t} \\ x_2(t) &= K_1 A_{112} e^{\lambda_1 t} \\ &\cdots \\ x_n(t) &= K_1 A_{11n} e^{\lambda_1 t} \end{aligned} \tag{3-14}$$

Clearly Eq. (3-14) defines the following hyper line (in the n-dimensional state space):

$$\frac{x_1}{A_{111}} = \frac{x_2}{A_{112}} = \cdots = \frac{x_n}{A_{11n}} \tag{3-15}$$

(Note that if $A_{11j} = 0$, then the corresponding x_j is also zero.) Equation (3-15) has been obtained for the eigenvalue λ_1. There are altogether n such hyper lines for n distinct eigenvalues λ_i ($i = 1, 2, 3, \ldots, n$). Thus, we get

$$\frac{x_1}{A_{i11}} = \frac{x_2}{A_{i12}} = \cdots = \frac{x_n}{A_{i1n}} \tag{3-16}$$

$$(i = 1, 2, \ldots, n)$$

Such hyper lines are eigenvectors. Eigenvectors are also called *principal coordinate axes*. [A_{kij} is the cofactor of a_{ij} of the determinant in Eq. (3-8) when λ_k is substituted for λ_i.] (Sometimes we may have $A_{i11} = A_{i12} = \cdots = A_{i1n} = 0$. We shall discuss this special case later.)

Notice that for real λ_i ($i = 1, 2, \ldots, n$) if the initial condition is on the eigenvector corresponding to λ_i, then the point representing the system state for $t > 0$ always lies on the same eigenvector. (If λ_i and λ_{i+1} are complex conjugate, however, then motions $e^{\lambda_i t}$ and $e^{\lambda_{i+1} t}$ cannot be separated. The combined motion takes place in the plane determined by the two eigenvectors corresponding to λ_i and λ_{i+1}. The motion is oscillatory. For detail, see Section 3-8.)

Coordinate transformation A point in the n-dimensional state space can be represented by a vector $\mathbf{x}$. If n eigenvectors of $\mathbf{A}$ are chosen as a new set of coordinate axes, the same point in the n-dimensional state space can be represented by a vector $\mathbf{y}$. Since the eigenvectors are given by Eq. (3-16), the relationship between the $\mathbf{x}$ coordinates and $\mathbf{y}$ coordinates can be given by the following equations:

$$\begin{aligned} x_1 &= A_{111} y_1 + A_{211} y_2 + \cdots + A_{n11} y_n \\ x_2 &= A_{112} y_1 + A_{212} y_2 + \cdots + A_{n12} y_n \\ &\cdots \qquad \cdots \qquad \cdots \\ x_n &= A_{11n} y_1 + A_{21n} y_2 + \cdots + A_{n1n} y_n \end{aligned} \tag{3-17}$$

provided we choose appropriate scales on coordinate axes. Equations (3-17)

can be rewritten in the following form:

$$\mathbf{x} = \mathbf{P}\mathbf{y}$$

where

$$\mathbf{P} = \begin{bmatrix} A_{111} & A_{211} & \cdots & A_{n11} \\ A_{112} & A_{212} & \cdots & A_{n12} \\ \vdots & \vdots & & \vdots \\ A_{11n} & A_{21n} & \cdots & A_{n1n} \end{bmatrix} \tag{3-18}$$

$\mathbf{P}$ is called a diagonalizing transformation matrix or simply transformation matrix. If this coordinate transformation is performed, we can show that Eq. (3-2) can be written as follows:

$$\dot{\mathbf{x}} = \mathbf{P}\dot{\mathbf{y}} = \mathbf{A}\mathbf{P}\mathbf{y}$$

or

$$\dot{\mathbf{y}} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}\mathbf{y} = \mathbf{D}\mathbf{y}$$

where $\mathbf{D}$ is a diagonal matrix. The initial condition terms are transformed into the following form:

$$\mathbf{y}(0) = \mathbf{P}^{-1}\mathbf{x}(0)$$

We shall now show the diagonalization of the coefficient matrix.

Let us compute $\mathbf{AP}$. By use of Eq. (3-9) $\mathbf{AP}$ can be obtained as follows:

$$\begin{aligned} \mathbf{AP} &= \begin{bmatrix} \lambda_1 A_{111} & \lambda_2 A_{211} & \cdots & \lambda_n A_{n11} \\ \lambda_1 A_{112} & \lambda_2 A_{212} & \cdots & \lambda_n A_{n12} \\ \vdots & \vdots & & \vdots \\ \lambda_1 A_{11n} & \lambda_2 A_{21n} & \cdots & \lambda_n A_{n1n} \end{bmatrix} \\ &= \begin{bmatrix} A_{111} & A_{211} & \cdots & A_{n11} \\ A_{112} & A_{212} & \cdots & A_{n12} \\ \vdots & \vdots & & \vdots \\ A_{11n} & A_{21n} & \cdots & A_{n1n} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} = \mathbf{PD} \end{aligned} \tag{3-19}$$

where

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

Since n columns of $\mathbf{P}$ are linearly independent, $\mathbf{P}$ is nonsingular and $\mathbf{P}^{-1}$ exists. Therefore

$$\mathbf{P}^{-1}\mathbf{AP} = \mathbf{D} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

With the transformation matrix $\mathbf{P}$ given by Eq. (3-18), Eq. (3-2) whose eigenvalues are distinct (which implies that the eigenvectors are also dis-

tinct) can be transformed into

$$\dot{\mathbf{y}} = \mathbf{D}\mathbf{y}$$

or

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \cdot \\ \cdot \\ \cdot \\ \dot{y}_n \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \tag{3–20}$$

This can be written in the following set of n first-order differential equations:

$$\dot{y}_1 = \lambda_1 y_1$$
$$\dot{y}_2 = \lambda_2 y_2$$
$$\cdot \quad \cdot \quad \cdot$$
$$\dot{y}_n = \lambda_n y_n$$

with initial conditions defined by

$$\mathbf{y}(0) = \mathbf{P}^{-1}\mathbf{x}(0)$$

Equation (3–20) shows that variables $y_1, y_2, \ldots, y_n$ are separated from each other.

The characteristic equation or eigenvalue equation of the system of Eq. (3–20) is

$$(\lambda_1 - \lambda)(\lambda_2 - \lambda)\cdots(\lambda_n - \lambda) = 0$$

There is no change in the eigenvalues. In fact, for any linear nonsingular transformation the eigenvalues remain the same.

Note that, if $\mathbf{P}$ is a diagonalizing transformation matrix, then matrix $\mathbf{PC}$ where $\mathbf{C}$ is a nonsingular diagonal matrix is also a diagonalizing transformation matrix, since if

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}$$

then

$$\mathbf{C}^{-1}\mathbf{P}^{-1}\mathbf{A}\mathbf{P}\mathbf{C} = \mathbf{C}^{-1}\mathbf{D}\mathbf{C} = \mathbf{D}$$

Hence, $\mathbf{P}$ defined by Eq. (3–18) is not the unique diagonalizing transformation matrix. Each column of $\mathbf{P}$ may be multiplied by a nonzero constant. In general, the transformation matrix can be written as

$$\begin{bmatrix} k_1 A_{111} & k_2 A_{211} & \cdots & k_n A_{n11} \\ k_1 A_{112} & k_2 A_{212} & \cdots & k_n A_{n12} \\ \vdots & \vdots & & \vdots \\ k_1 A_{11n} & k_2 A_{21n} & \cdots & k_n A_{n1n} \end{bmatrix} \tag{3–21}$$

where $k_1, k_2, \ldots, k_n$ are nonzero constants. (Notice that each column represents an eigenvector. Eigenvectors determine only directions of the coordinate axes in the n-dimensional state space.)

Special case where $A_{i11} = A_{i12} = \cdots = A_{i1n} = 0$ Sometimes it happens that $A_{i11} = A_{i12} = \cdots = A_{i1n} = 0$. In this case, we may take $A_{ik1}, A_{ik2}, \ldots, A_{ikn}$ $(k \neq 1)$ to form the ith eigenvector. [Notice that in obtaining Eq. (3–9) we expanded the determinant of Eq. (3–8) about the first row. This, however, is not necessary. We can expand the determinant about the kth row $(k \neq 1)$ as well.] Consider the following example:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 2 & -1 & 0 \\ 5 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

The three eigenvalues of the coefficient matrix are $\lambda_1 = -2$, $\lambda_2 = -1$, $\lambda_3 = 1$. Let us compute the matrix defined by Eq. (3–18) for this example:

$$\begin{bmatrix} 1 & 0 & 4 \\ -2 & 0 & 4 \\ -5 & 0 & 10 \end{bmatrix}$$

The elements of the second column are zeros. Hence the inverse of this matrix does not exist. Instead of $A_{211}, A_{212}, A_{213}$, let us take $A_{231}, A_{232}, A_{233}$ as the elements of the second column. (The elements of the first and third columns remain the same.) Then,

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 4 \\ -2 & 0 & 4 \\ -5 & -2 & 10 \end{bmatrix}$$

and

$$\mathbf{P}^{-1} = \begin{bmatrix} \frac{1}{3} & -\frac{1}{3} & 0 \\ 0 & \frac{5}{4} & -\frac{1}{2} \\ \frac{1}{6} & \frac{1}{12} & 0 \end{bmatrix}$$

The matrix $\mathbf{P}$ obtained in this way will transform $\mathbf{A}$ into the diagonal form. To check, let us compute $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$.

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{bmatrix} \frac{1}{3} & -\frac{1}{3} & 0 \\ 0 & \frac{5}{4} & -\frac{1}{2} \\ \frac{1}{6} & \frac{1}{12} & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 2 & -1 & 0 \\ 5 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 4 \\ -2 & 0 & 4 \\ -5 & -2 & 10 \end{bmatrix}$$
$$= \begin{bmatrix} -2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Computation of diagonalizing transformation matrix $\mathbf{P}$ (Applicable when A has distinct eigenvalues) Note that, in free dynamic systems with distinct eigenvalues, once the distinct eigenvalues of the system are

known it is not necessary to find the matrix $\mathbf{P}$ in order to obtain the diagonal form. The diagonal form can be written directly from the knowledge of the eigenvalues. In forced dynamic systems, however, we need $\mathbf{P}^{-1}$. (Refer to Chapter 4.) In the following we shall discuss two methods for finding $\mathbf{P}$.

It should be noted that if $\mathbf{A}$ has one of the following two forms:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_2 & -a_1 \end{bmatrix}$$

or

$$\mathbf{A} = \begin{bmatrix} -a_1 & 1 & 0 & \cdots & 0 & 0 \\ -a_2 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ -a_{n-1} & 0 & 0 & \cdots & 0 & 1 \\ -a_n & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

then simple formulas are available for obtaining a diagonalizing transformation matrix $\mathbf{P}$. [Refer to Eq. (3–37), Eq. (3–50), and Eq. (3–52); and Prob. A-3-10 and Prob. A-3-11.]

Method 1 A diagonalizing transformation matrix $\mathbf{P}$ can be obtained by computing A_{ij} directly. For example, consider the following matrix:

$$\mathbf{A} = \begin{bmatrix} 4 & 2 \\ -3 & -1 \end{bmatrix}$$

The eigenvalues of $\mathbf{A}$ are 1 and 2. Matrix $\mathbf{P}$ is given by

$$\mathbf{P} = \begin{bmatrix} A_{111} & A_{211} \\ A_{112} & A_{212} \end{bmatrix}$$

In order to find A_{111} and A_{112} let us substitute $\lambda_1 = 1$ in the characteristic polynomial. Then we obtain

$$|\mathbf{A} - \lambda_1\mathbf{I}| = \begin{vmatrix} 3 & 2 \\ -3 & -2 \end{vmatrix}$$

Hence,

$$A_{111} = -2, \qquad A_{112} = -(-3)$$

To find A_{211} and A_{212} let us substitute $\lambda_2 = 2$ in the characteristic polynomial.

$$|\mathbf{A} - \lambda_2\mathbf{I}| = \begin{vmatrix} 2 & 2 \\ -3 & -3 \end{vmatrix}$$

We thus obtain

$$A_{211} = -3, \qquad A_{212} = -(-3)$$

We therefore find $\mathbf{P}$ as follows:

$$\mathbf{P} = \begin{bmatrix} -2 & -3 \\ 3 & 3 \end{bmatrix}$$

Then we obtain

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{bmatrix} 1 & 1 \\ -1 & -\frac{2}{3} \end{bmatrix}\begin{bmatrix} 4 & 2 \\ -3 & -1 \end{bmatrix}\begin{bmatrix} -2 & -3 \\ 3 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

Formulas for finding diagonalizing transformation matrices $\mathbf{P}$ that transform the 2×2 constant matrix and 3×3 constant matrix into the diagonal form, respectively, are given in the following:

(1) For the following 2×2 constant matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

where $|\mathbf{A} - \lambda\mathbf{I}| = (\lambda - \lambda_1)(\lambda - \lambda_2), \qquad \lambda_1 \neq \lambda_2$

a diagonalizing transformation matrix $\mathbf{P}$ that transforms $\mathbf{A}$ into a diagonal matrix is given by

$$\mathbf{P} = \begin{bmatrix} A_{111} & A_{211} \\ A_{112} & A_{212} \end{bmatrix} = \begin{bmatrix} d - \lambda_1 & d - \lambda_2 \\ -c & -c \end{bmatrix}$$

(2) For the following 3×3 constant matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

where $|\mathbf{A} - \lambda\mathbf{I}| = -(\lambda - \lambda_1)(\lambda - \lambda_2)(\lambda - \lambda_3)$

and λ_1, λ_2, and λ_3 are different from each other, a diagonalizing transformation matrix $\mathbf{P}$ that transforms $\mathbf{A}$ into a diagonal matrix is given by

$$\mathbf{P} = \begin{bmatrix} A_{111} & A_{211} & A_{311} \\ A_{112} & A_{212} & A_{312} \\ A_{113} & A_{213} & A_{313} \end{bmatrix}$$

$$= \begin{bmatrix} \begin{vmatrix} e - \lambda_1 & f \\ h & i - \lambda_1 \end{vmatrix} & \begin{vmatrix} e - \lambda_2 & f \\ h & i - \lambda_2 \end{vmatrix} & \begin{vmatrix} e - \lambda_3 & f \\ h & i - \lambda_3 \end{vmatrix} \\ -\begin{vmatrix} d & f \\ g & i - \lambda_1 \end{vmatrix} & -\begin{vmatrix} d & f \\ g & i - \lambda_2 \end{vmatrix} & -\begin{vmatrix} d & f \\ g & i - \lambda_3 \end{vmatrix} \\ \begin{vmatrix} d & e - \lambda_1 \\ g & h \end{vmatrix} & \begin{vmatrix} d & e - \lambda_2 \\ g & h \end{vmatrix} & \begin{vmatrix} d & e - \lambda_3 \\ g & h \end{vmatrix} \end{bmatrix}$$

Method 2 Let us write

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}$$

Since each column of $\mathbf{P}$ is an eigenvector of $\mathbf{A}$ we have

$$\mathbf{AP}_i = \lambda_i \mathbf{P}_i$$

where $\mathbf{P}_i$ is the ith column of the matrix $\mathbf{P}$. To compute the elements of the column matrix $\mathbf{P}_i$, we solve

$$(\mathbf{A} - \lambda_i \mathbf{I})\mathbf{P}_i = \mathbf{0} \qquad (i = 1, 2, \ldots, n)$$

for $p_{1i}, p_{2i}, \ldots, p_{ni}$.

For example, consider the same matrix $\mathbf{A}$ as in the example of Method 1, or

$$\mathbf{A} = \begin{bmatrix} 4 & 2 \\ -3 & -1 \end{bmatrix}$$

The eigenvalues of $\mathbf{A}$ are 1 and 2. To obtain the diagonalizing transformation matrix $\mathbf{P}$, we solve the two equations

$$(\mathbf{A} - \mathbf{I})\mathbf{P}_1 = \mathbf{0} \qquad \text{and} \qquad (\mathbf{A} - 2\mathbf{I})\mathbf{P}_2 = \mathbf{0}$$

which can be written as

$$\begin{bmatrix} 3 & 2 \\ -3 & -2 \end{bmatrix}\begin{bmatrix} p_{11} \\ p_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} 2 & 2 \\ -3 & -3 \end{bmatrix}\begin{bmatrix} p_{12} \\ p_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

These two matrix equations yield the following two scalar equations:

$$3p_{11} + 2p_{21} = 0$$
$$p_{12} + p_{22} = 0$$

Since we have four unknowns and two equations, we can choose two unknowns arbitrarily. Let us choose $p_{11} = 1$ and $p_{12} = 1$. We then obtain

$$p_{21} = -\tfrac{3}{2}, \qquad p_{22} = -1$$

Thus

$$\mathbf{P} = \begin{bmatrix} 1 & 1 \\ -\frac{3}{2} & -1 \end{bmatrix}$$

and

$$\mathbf{P}^{-1}\mathbf{AP} = \begin{bmatrix} -2 & -2 \\ 3 & 2 \end{bmatrix}\begin{bmatrix} 4 & 2 \\ -3 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ -\frac{3}{2} & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

System whose distinct eigenvalues involve complex conjugate pairs Assume for convenience that the system involves only one pair of complex conjugate eigenvalues. This, however, does not limit the usefulness of the following discussion in any way, since extension to the case where there are more than one pair of complex conjugate eigenvalues is obvious.

Assume that the eigenvalues λ_1 and $\bar{\lambda}_1$ are complex conjugates and are given by

$$\lambda_1 = \sigma + j\omega, \qquad \bar{\lambda}_1 = \sigma - j\omega$$

Assume also that the eigenvalues $\lambda_3, \lambda_4, \ldots, \lambda_n$ are real and distinct. We can transform Eq. (3–2) into the form given by Eq. (3–20) or

$$\dot{y}_1 = (\sigma + j\omega)y_1$$
$$\dot{y}_2 = (\sigma - j\omega)y_2$$
$$\dot{y}_3 = \lambda_3 y_3$$
$$\cdots$$
$$\dot{y}_n = \lambda_n y_n$$

If the second transformation is made by the following equations:

$$y_1 = \frac{1}{2}z_1 - \frac{j}{2}z_2$$
$$y_2 = \frac{1}{2}z_1 + \frac{j}{2}z_2$$
$$y_3 = z_3$$
$$\cdots$$
$$y_n = z_n$$

or

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -\frac{j}{2} & 0 & \cdots & 0 \\ \frac{1}{2} & \frac{j}{2} & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_n \end{bmatrix}$$

then we obtain

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \\ \vdots \\ \dot{z}_n \end{bmatrix} = \begin{bmatrix} \sigma & \omega & 0 & \cdots & 0 \\ -\omega & \sigma & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_n \end{bmatrix}$$

or

$$\dot{z}_1 = \sigma z_1 + \omega z_2$$
$$\dot{z}_2 = -\omega z_1 + \sigma z_2$$
$$\dot{z}_3 = \lambda_3 z_3$$
$$\cdots$$
$$\dot{z}_n = \lambda_n z_n$$

It can be seen that z_1 and z_2 are dependent on each other, although z_1 (or z_2), $z_3, z_4, \ldots, z_n$ are independent of each other.

If two pairs of conjugate eigenvalues appear in the characteristic equation, then the system equation can be given as follows:

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \\ \dot{z}_4 \\ \dot{z}_5 \\ \vdots \\ \dot{z}_n \end{bmatrix} = \begin{bmatrix} \sigma_1 & \omega_1 & & & & & 0 \\ -\omega_1 & \sigma_1 & & & & & \\ & & \sigma_2 & \omega_2 & & & \\ & & -\omega_2 & \sigma_2 & & & \\ & & & & \lambda_5 & & 0 \\ & & & & & \ddots & \\ 0 & & & & 0 & & \lambda_n \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ \vdots \\ z_n \end{bmatrix}$$

This completes our preliminary discussions on diagonalization of coefficient matrices in state space equations which are vector matrix differential equations. In Section 3–3 we discuss details of diagonalization of various types of matrices similar to diagonal matrices.

3-3 DIAGONALIZATION OF MATRICES

In this section we present the necessary and sufficient condition that a given matrix is similar to a diagonal matrix. We shall also present associated subjects on diagonalization of matrices in the form of theorems and proofs.

We shall first prove the following basic theorem:

Theorem 3-1 An $n \times n$ matrix $\mathbf{A}$ is similar to a diagonal matrix if and only if a set of n eigenvectors is linearly independent.

Proof Suppose that $\mathbf{A}$ has n linearly independent eigenvectors $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_n$. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the corresponding eigenvalues. Let us define

$$\mathbf{P} = [\mathbf{P}_1 \mid \mathbf{P}_2 \mid \cdots \mid \mathbf{P}_n]$$

Since

$$\mathbf{A}\mathbf{P}_i = \lambda_i \mathbf{P}_i \qquad (i = 1, 2, \ldots, n)$$

we have

$$\mathbf{AP} = \mathbf{PD}$$

where

$$\mathbf{D} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

Since n column vectors $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_n$ of $\mathbf{P}$ are linearly independent, $\mathbf{P}$ is nonsingular, and therefore $\mathbf{P}^{-1}$ exists. Hence we obtain

$$\mathbf{P}^{-1}\mathbf{AP} = \mathbf{D}$$

It follows that if n eigenvectors are linearly independent, $\mathbf{A}$ is similar to a diagonal matrix.

Conversely, if $\mathbf{A}$ is similar to a diagonal matrix, then there exists a nonsingular matrix $\mathbf{Q}$ such that

$$\mathbf{Q}^{-1}\mathbf{AQ} = \mathbf{D} \qquad \text{or} \qquad \mathbf{AQ} = \mathbf{QD}$$

Let

$$\mathbf{Q} = [\mathbf{Q}_1 \mid \mathbf{Q}_2 \mid \cdots \mid \mathbf{Q}_n]$$

Then

$$\mathbf{A}[\mathbf{Q}_1 \mid \mathbf{Q}_2 \mid \cdots \mid \mathbf{Q}_n] = [\mathbf{Q}_1 \mid \mathbf{Q}_2 \mid \cdots \mid \mathbf{Q}_n]\mathbf{D}$$

or

$$\mathbf{A}\mathbf{Q}_i = \lambda_i \mathbf{Q}_i \qquad (i = 1, 2, \ldots, n)$$

This implies that the $\mathbf{Q}_i$ are n eigenvectors of $\mathbf{A}$. Since $\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_n$ are column vectors of a nonsingular matrix $\mathbf{Q}$, they are linearly independent. Thus we have proved that $\mathbf{A}$ is similar to a diagonal matrix if and only if n eigenvectors are linearly independent.

Notice that if an $n \times n$ matrix $\mathbf{A}$ having n distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ is similar to a diagonal matrix $\mathbf{D}$, then the diagonal elements of $\mathbf{D}$ are the eigenvalues of $\mathbf{A}$. Hence the diagonal matrix $\mathbf{D}$ is unique, except the order of the diagonal elements $\lambda_1, \lambda_2, \ldots, \lambda_n$.

We shall next prove the following theorem:

Theorem 3-2 An $n \times n$ matrix $\mathbf{A}$ possessing n distinct eigenvalues has n linearly independent eigenvectors.

Proof Let n eigenvalues of $\mathbf{A}$ be $\lambda_1, \lambda_2, \ldots, \lambda_n$ and corresponding eigenvectors be $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_n$, respectively. Then we have

$$\mathbf{A}\mathbf{P}_i = \lambda_i \mathbf{P}_i \qquad (i = 1, 2, \ldots, n)$$

Consider the equation

$$c_1\mathbf{P}_1 + c_2\mathbf{P}_2 + \cdots + c_n\mathbf{P}_n = 0$$

If n eigenvectors $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_n$ are linearly independent, then $c_1, c_2, \ldots, c_n$ must be equal to zero. Multiplying

$$(\mathbf{A} - \lambda_1\mathbf{I}) \cdots (\mathbf{A} - \lambda_{i-1}\mathbf{I})(\mathbf{A} - \lambda_{i+1}\mathbf{I}) \cdots (\mathbf{A} - \lambda_n\mathbf{I})$$

into the foregoing equation we obtain

$$\begin{aligned}(\mathbf{A} - \lambda_1\mathbf{I}) \cdots (\mathbf{A} - \lambda_{i-1}\mathbf{I})(\mathbf{A} - \lambda_{i+1}\mathbf{I}) \cdots (\mathbf{A} - \lambda_n\mathbf{I}) \\ \cdot(c_1\mathbf{P}_1 + c_2\mathbf{P}_2 + \cdots + c_i\mathbf{P}_i + \cdots + c_n\mathbf{P}_n) \\ = (\mathbf{A} - \lambda_1\mathbf{I}) \cdots (\mathbf{A} - \lambda_{i-1}\mathbf{I})(\mathbf{A} - \lambda_{i+1}\mathbf{I}) \cdots (\mathbf{A} - \lambda_n\mathbf{I})c_i\mathbf{P}_i \\ = (\lambda_i - \lambda_1) \cdots (\lambda_i - \lambda_{i-1})(\lambda_i - \lambda_{i+1}) \cdots (\lambda_i - \lambda_n)c_i\mathbf{P}_i = \mathbf{0} \qquad (3\text{–}22)\end{aligned}$$

where we have used the fact that

$$(\mathbf{A} - \lambda_j\mathbf{I})(\mathbf{A} - \lambda_k\mathbf{I}) = (\mathbf{A} - \lambda_k\mathbf{I})(\mathbf{A} - \lambda_j\mathbf{I})$$

and

$$\begin{aligned}(\mathbf{A} - \lambda_j\mathbf{I})c_k\mathbf{P}_k &= \mathbf{A}c_k\mathbf{P}_k - \lambda_j\mathbf{I}c_k\mathbf{P}_k \\ &= \lambda_k c_k\mathbf{P}_k - \lambda_j c_k\mathbf{P}_k \\ &= (\lambda_k - \lambda_j)c_k\mathbf{P}_k \\ &= \mathbf{0} \qquad (k = j) \\ &\neq \mathbf{0} \qquad (k \neq j)\end{aligned}$$

Since the λ_i are distinct, Eq. (3–22) implies that $c_i = 0$ $(i = 1, 2, \ldots, n)$. Thus, the n eigenvectors corresponding to n distinct eigenvalues are linearly independent. This completes the proof.

The following theorem follows immediately from Theorems 3–1 and 3–2:

Theorem 3-3 An $n \times n$ matrix $\mathbf{A}$ with distinct eigenvalues is similar to a diagonal matrix.

From Theorems 3–1, 3–2, and 3–3, we deduce that if an $n \times n$ matrix $\mathbf{A}$ is to be similar to a diagonal matrix, then for each multiple eigenvalue λ_i of multiplicity m_i there must exist m_i linearly independent eigenvectors corresponding to λ_i.

We shall next prove that any $n \times n$ matrix can be transformed into a triangular matrix.

Theorem 3-4 An $n \times n$ matrix $\mathbf{A}$, whose eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are not necessarily distinct, can always be reduced to a triangular form, or

$$\mathbf{Z}^{-1}\mathbf{A}\mathbf{Z} = \begin{bmatrix} \lambda_1 & * & & * \\ & \lambda_2 & & \\ & & \ddots & \\ & & & * \\ 0 & & & \lambda_n \end{bmatrix}$$

where $\mathbf{Z}$ is a nonsingular matrix and the $*$ denote elements not necessarily zero.

Proof We shall first prove that there exists a nonsingular matrix $\mathbf{X}_1$ such that

$$\mathbf{X}_1^{-1}\mathbf{A}\mathbf{X}_1 = \left[\begin{array}{c|c} \lambda_1 & \mathbf{a}_1 \\ \hline \mathbf{0} & \mathbf{A}_1 \end{array}\right]$$

where
$\mathbf{A}_1 = (n-1) \times (n-1)$ matrix
$\mathbf{a}_1 = 1 \times (n-1)$ matrix
$\mathbf{0}$ = matrix consisting of zeros

Let us denote the nonzero solution of

$$\mathbf{A}\mathbf{x} = \lambda_1\mathbf{x}$$

as $\mathbf{x}_1$. (Such solution $\mathbf{x}_1$ exists, since $|\mathbf{A} - \lambda_1\mathbf{I}| = 0$.) Let us define $\mathbf{X}_1$ as a nonsingular matrix having $\mathbf{x}_1$ as its first column. Let us denote n columns of $\mathbf{A}\mathbf{X}_1$ as $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_n$, or

$$\mathbf{A}\mathbf{X}_1 = [\mathbf{P}_1 \mid \mathbf{P}_2 \mid \cdots \mid \mathbf{P}_n]$$

and n columns of $\mathbf{X}_1^{-1}\mathbf{A}\mathbf{X}_1$ as $\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_n$, or

$$\mathbf{X}_1^{-1}\mathbf{A}\mathbf{X}_1 = [\mathbf{Q}_1 \mid \mathbf{Q}_2 \mid \cdots \mid \mathbf{Q}_n]$$

Then clearly

$$\mathbf{P}_1 = \mathbf{A}\mathbf{x}_1 = \lambda_1\mathbf{x}_1$$

Since

$$\mathbf{X}_1^{-1} = \frac{1}{|\mathbf{X}_1|}\begin{bmatrix} \Delta_{11} & \Delta_{21} & \cdots & \Delta_{n1} \\ \Delta_{12} & \Delta_{22} & \cdots & \Delta_{n2} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \Delta_{1n} & \Delta_{2n} & \cdots & \Delta_{nn} \end{bmatrix}$$

where Δ_{ij} is the cofactor of the element x_{ij}, we obtain

$$\mathbf{Q}_1 = \frac{1}{|\mathbf{X}_1|}\begin{bmatrix} \lambda_1 \sum_{j=1}^{n} \Delta_{j1}x_{j1} \\ \lambda_1 \sum_{j=1}^{n} \Delta_{j2}x_{j1} \\ \cdot \\ \cdot \\ \cdot \\ \lambda_1 \sum_{j=1}^{n} \Delta_{jn}x_{j1} \end{bmatrix} = \frac{1}{|\mathbf{X}_1|}\begin{bmatrix} \lambda_1 |\mathbf{X}_1| \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

Thus, we have shown that the matrix $\mathbf{X}_1^{-1}\mathbf{A}\mathbf{X}_1$, where $\mathbf{X}_1$ has $\mathbf{x}_1$ as its first column, will have λ_1 as the first element of the first column and zeros in the rest of the first column.

The matrix $\mathbf{A}_1$ has $n-1$ eigenvalues, $\lambda_2, \lambda_3, \ldots, \lambda_n$. Let us define the solution of $\mathbf{A}_1\mathbf{y} = \lambda_2\mathbf{y}$ as $\mathbf{y}_1$. By choosing $\mathbf{y}_1$ as the first column of a nonsingular matrix $\mathbf{Y}$, we have

$$\mathbf{Y}^{-1}\mathbf{A}_1\mathbf{Y} = \left[\begin{array}{c|c} \lambda_2 & \mathbf{a}_2 \\ \hline \mathbf{0} & \mathbf{A}_2 \end{array}\right]$$

where $\mathbf{A}_2 = (n-2) \times (n-2)$ matrix

$\mathbf{a}_2 = 1 \times (n-2)$ matrix

Notice that by choosing $\mathbf{X}_2$ as

$$\mathbf{X}_2 = \left[\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{Y} \end{array}\right]$$

we obtain

$$\mathbf{X}_2^{-1}(\mathbf{X}_1^{-1}\mathbf{A}\mathbf{X}_1)\mathbf{X}_2 = \left[\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{Y} \end{array}\right]^{-1} \left[\begin{array}{c|c} \lambda_1 & \mathbf{a}_1 \\ \hline \mathbf{0} & \mathbf{A}_1 \end{array}\right] \left[\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{Y} \end{array}\right]$$

$$= \left[\begin{array}{c|c} \lambda_1 & \mathbf{a}_1\mathbf{Y} \\ \hline \mathbf{0} & \mathbf{Y}^{-1}\mathbf{A}_1\mathbf{Y} \end{array}\right] = \left[\begin{array}{cc|c} \lambda_1 & * & * \\ 0 & \lambda_2 & \\ \hline & \mathbf{0} & \mathbf{A}_2 \end{array}\right]$$

where the $*$ denote elements not necessarily zero. Repeating the same process, we finally obtain

$$\mathbf{Z}^{-1}\mathbf{A}\mathbf{Z} = \begin{bmatrix} \lambda_1 & * & & * \\ & \lambda_2 & & \\ & & \ddots & \\ & & & * \\ 0 & & & \lambda_n \end{bmatrix}$$

where $\mathbf{Z} = \mathbf{X}_1\mathbf{X}_2 \cdots \mathbf{X}_n$.

This completes the proof.

Notice that Theorem 3–4 is useful in solving vector matrix differential equations. Consider the vector matrix differential equation

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}, \qquad \mathbf{x}(0) = \mathbf{x}_0$$

This equation can be transformed, by use of the transformation $\mathbf{x} = \mathbf{Z}\mathbf{y}$, into

$$\dot{\mathbf{y}} = \mathbf{Z}^{-1}\mathbf{A}\mathbf{Z}\mathbf{y}, \qquad \mathbf{y}(0) = \mathbf{Z}^{-1}\mathbf{x}(0)$$

Since $\mathbf{Z}^{-1}\mathbf{A}\mathbf{Z}$ is a triangular matrix this last differential equation can be solved for $\mathbf{y}$ by solving first the scalar differential equation in y_n and then substituting the solution y_n into the scalar differential equation in y_{n-1}, then substituting the solutions y_n and y_{n-1} into the scalar differential equation in y_{n-2}. Repeating this process, $y_n, y_{n-1}, \ldots, y_1$ can be obtained. By the transformation $\mathbf{x} = \mathbf{Z}\mathbf{y}$, we can obtain the solution $\mathbf{x}$.

Unitarily similar matrices An $n \times n$ matrix $\mathbf{B}$ is called *unitarily similar* to an $n \times n$ matrix $\mathbf{A}$ if there exists a unitary matrix $\mathbf{U}$ such that $\mathbf{B} = \mathbf{U}^{-1}\mathbf{A}\mathbf{U}$.

Theorem 3-5 An $n \times n$ matrix $\mathbf{A}$ over the field of complex numbers is unitarily similar to a diagonal matrix if and only if it is normal, or

$$\mathbf{A}\mathbf{A}^* = \mathbf{A}^*\mathbf{A}$$

Proof Suppose that $\mathbf{A}$ is normal. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of $\mathbf{A}$ and $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be normalized eigenvectors associated with $\lambda_1, \lambda_2, \ldots, \lambda_n$, respectively. Let $\mathbf{U}_1$ be a unitary matrix. If the first column of $\mathbf{U}$ is chosen to be $\mathbf{x}_1$, then refering to Theorem 3–4 the first column of $\mathbf{U}_1^{-1}\mathbf{A}\mathbf{U}_1$ becomes equal to

$$\begin{bmatrix} \lambda_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Hence we may write

$$\mathbf{U}_1^{-1}\mathbf{A}\mathbf{U}_1 = \left[\begin{array}{c|c} \lambda_1 & \mathbf{a}_1 \\ \hline \mathbf{0} & \mathbf{A}_1 \end{array}\right]$$

where $\mathbf{A}_1 = (n-1) \times (n-1)$ normal matrix whose eigenvalues are $\lambda_2, \lambda_3, \ldots, \lambda_n$

$\mathbf{a}_1 = 1 \times (n-1)$ matrix

$\mathbf{0}$ = matrix consisting of zeros

Since

$$\begin{aligned}(\mathbf{U}_1^{-1}\mathbf{A}\mathbf{U}_1)(\mathbf{U}_1^{-1}\mathbf{A}\mathbf{U}_1)^* &= \mathbf{U}_1^*\mathbf{A}\mathbf{U}_1\mathbf{U}_1^*\mathbf{A}^*\mathbf{U}_1 \\ &= \mathbf{U}_1^*\mathbf{A}\mathbf{A}^*\mathbf{U}_1 = \mathbf{U}_1^*\mathbf{A}^*\mathbf{A}\mathbf{U}_1 \\ &= \mathbf{U}_1^*\mathbf{A}^*(\mathbf{U}_1^{-1})^*\mathbf{U}_1^{-1}\mathbf{A}\mathbf{U}_1 = (\mathbf{U}_1^{-1}\mathbf{A}\mathbf{U}_1)^*(\mathbf{U}_1^{-1}\mathbf{A}\mathbf{U}_1)\end{aligned}$$

$\mathbf{U}_1^{-1}\mathbf{A}\mathbf{U}_1$ is normal. Hence

$$\left[\begin{array}{c|c} \lambda_1 & \mathbf{a}_1 \\ \hline \mathbf{0} & \mathbf{A}_1 \end{array}\right] \left[\begin{array}{c|c} \bar{\lambda}_1 & \mathbf{0} \\ \hline \mathbf{a}_1^* & \mathbf{A}_1^* \end{array}\right] = \left[\begin{array}{c|c} \bar{\lambda}_1 & \mathbf{0} \\ \hline \mathbf{a}_1^* & \mathbf{A}_1^* \end{array}\right] \left[\begin{array}{c|c} \lambda_1 & \mathbf{a}_1 \\ \hline \mathbf{0} & \mathbf{A}_1 \end{array}\right]$$

or

$$\left[\begin{array}{c|c} \lambda_1\bar{\lambda}_1 + \mathbf{a}_1\mathbf{a}_1^* & \mathbf{a}_1\mathbf{A}_1^* \\ \hline \mathbf{A}_1\mathbf{a}_1^* & \mathbf{A}_1\mathbf{A}_1^* \end{array}\right] = \left[\begin{array}{c|c} \bar{\lambda}_1\lambda_1 & \bar{\lambda}_1\mathbf{a}_1 \\ \hline \mathbf{a}_1^*\lambda_1 & \mathbf{a}_1^*\mathbf{a}_1 + \mathbf{A}_1^*\mathbf{A}_1 \end{array}\right]$$

it follows that $\mathbf{a}_1$ must be equal to zero, or $\mathbf{a}_1 = \mathbf{0}$. Hence $\mathbf{U}_1^{-1}\mathbf{A}\mathbf{U}_1$ can be rewritten as

$$\mathbf{U}_1^{-1}\mathbf{A}\mathbf{U}_1 = \left[\begin{array}{c|c} \lambda_1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{A}_1 \end{array}\right]$$

where $\mathbf{A}_1$ is an $(n-1) \times (n-1)$ normal matrix whose eigenvalues are $\lambda_2, \lambda_3, \ldots, \lambda_n$.

By repeating the same process, we can find an $(n-1) \times (n-1)$ unitary matrix $\mathbf{V}$ such that

$$\mathbf{V}^{-1}\mathbf{A}_1\mathbf{V} = \left[\begin{array}{c|c} \lambda_2 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{A}_2 \end{array}\right]$$

where $\mathbf{A}_2$ is an $(n-2) \times (n-2)$ normal matrix whose eigenvalues are $\lambda_3, \lambda_4, \ldots, \lambda_n$. By letting

$$\mathbf{U}_2 = \left[\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{V} \end{array}\right]$$

we obtain

$$\mathbf{U}_2^{-1}(\mathbf{U}_1^{-1}\mathbf{A}\mathbf{U}_1)\mathbf{U}_2 = \left[\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{V}^{-1} \end{array}\right]\left[\begin{array}{c|c} \lambda_1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{A}_1 \end{array}\right]\left[\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{V} \end{array}\right]$$

$$= \left[\begin{array}{c|c} \lambda_1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{V}^{-1}\mathbf{A}_1\mathbf{V} \end{array}\right] = \left[\begin{array}{cc|c} \lambda_1 & 0 & \mathbf{0} \\ 0 & \lambda_2 & \\ \hline & \mathbf{0} & \mathbf{A}_2 \end{array}\right]$$

Repeating the same process, we can finally obtain

$$\mathbf{U}^{-1}\mathbf{A}\mathbf{U} = \mathbf{U}^*\mathbf{A}\mathbf{U} = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix}$$

where

$$\mathbf{U} = \mathbf{U}_1\mathbf{U}_2 \cdots \mathbf{U}_n$$

Hence we have proved that $\mathbf{A}$ is unitarily similar to a diagonal matrix if $\mathbf{A}$ is normal.

We shall next prove that if $\mathbf{A}$ is unitarily similar to a diagonal matrix, then $\mathbf{A}$ is normal. Let

$$\mathbf{U}^{-1}\mathbf{A}\mathbf{U} = \mathbf{D} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

where $\mathbf{U}$ is unitary. Then,

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^{-1}$$

and

$$\begin{aligned} \mathbf{A}\mathbf{A}^* &= (\mathbf{U}\mathbf{D}\mathbf{U}^{-1})(\mathbf{U}\mathbf{D}\mathbf{U}^{-1})^* = \mathbf{U}\mathbf{D}\mathbf{U}^{-1}\mathbf{U}\mathbf{D}^*\mathbf{U}^* \\ &= \mathbf{U}\mathbf{D}\mathbf{D}^*\mathbf{U}^* = \mathbf{U}\mathbf{D}^*\mathbf{D}\mathbf{U}^* \\ &= (\mathbf{U}^{-1})^*\mathbf{D}^*\mathbf{U}^*\mathbf{U}\mathbf{D}\mathbf{U}^{-1} = (\mathbf{U}\mathbf{D}\mathbf{U}^{-1})^*(\mathbf{U}\mathbf{D}\mathbf{U}^{-1}) \\ &= \mathbf{A}^*\mathbf{A} \end{aligned}$$

Thus we have proved that $\mathbf{A}$ is normal. Hence the theorem.

From Theorem 3-5 we deduce that Hermitian matrices, skew Hermitian matrices, unitary matrices, and orthogonal matrices are unitarily similar to diagonal matrices. Hence these matrices can be diagonalized.

Example 3-1 Diagonalize the coefficient matrix of the following state space equation:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} 1 & 1+j\sqrt{2} & 0 \\ 1-j\sqrt{2} & 2 & j\sqrt{3} \\ 0 & -j\sqrt{3} & 1 \end{bmatrix}$$

Notice that $\mathbf{A}$ is Hermitian; therefore it can be diagonalized by use of the transformation matrix whose columns consist of the normalized eigenvectors of $\mathbf{A}$.

The eigenvalues of $\mathbf{A}$ are found to be

$$\lambda_1 = 1, \qquad \lambda_2 = -1, \qquad \lambda_3 = 4$$

For $\lambda_1 = 1$, the eigenvector is obtained from solving $(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{P}_1 = \mathbf{0}$ for $\mathbf{P}_1$. Namely the eigenvector $\mathbf{P}_1$ is determined from

$$\begin{aligned} (1+j\sqrt{2})\,p_{21} &= 0 \\ (1-j\sqrt{2})\,p_{11} + p_{21} + j\sqrt{3}\,p_{31} &= 0 \\ -j\sqrt{3}\,p_{21} &= 0 \end{aligned}$$

it follows that

$$\mathbf{P}_1 = \begin{bmatrix} p_{11} \\ p_{21} \\ p_{31} \end{bmatrix} = \begin{bmatrix} -j\sqrt{3}\,a \\ 0 \\ (1-j\sqrt{2})\,a \end{bmatrix}, \qquad a = \text{nonzero constant}$$

Similarly for $\lambda_2 = -1$, we obtain

$$\mathbf{P}_2 = \begin{bmatrix} p_{12} \\ p_{22} \\ p_{32} \end{bmatrix} = \begin{bmatrix} (1+j\sqrt{2})\,b \\ -2b \\ -j\sqrt{3}\,b \end{bmatrix}, \qquad b = \text{nonzero constant}$$

and for $\lambda_3 = 4$ we obtain

$$\mathbf{P}_3 = \begin{bmatrix} p_{13} \\ p_{23} \\ p_{33} \end{bmatrix} = \begin{bmatrix} (1+j\sqrt{2})\,c \\ 3c \\ -j\sqrt{3}\,c \end{bmatrix}, \qquad c = \text{nonzero constant}$$

Hence

$$\mathbf{P} = [\mathbf{P}_1 \,|\, \mathbf{P}_2 \,|\, \mathbf{P}_3] = \begin{bmatrix} -j\sqrt{3}\,a & (1+j\sqrt{2})\,b & (1+j\sqrt{2})\,c \\ 0 & -2b & 3c \\ (1-j\sqrt{2})\,a & -j\sqrt{3}\,b & -j\sqrt{3}\,c \end{bmatrix}$$

By normalizing each column vector, we obtain the following unitary matrix $\mathbf{U}$:

$$\mathbf{U} = \begin{bmatrix} \dfrac{-j\sqrt{3}}{\sqrt{6}} & \dfrac{1+j\sqrt{2}}{\sqrt{10}} & \dfrac{1+j\sqrt{2}}{\sqrt{15}} \\ 0 & \dfrac{-2}{\sqrt{10}} & \dfrac{3}{\sqrt{15}} \\ \dfrac{1-j\sqrt{2}}{\sqrt{6}} & \dfrac{-j\sqrt{3}}{\sqrt{10}} & \dfrac{-j\sqrt{3}}{\sqrt{15}} \end{bmatrix}$$

The unitary transformation $\mathbf{x} = \mathbf{U}\mathbf{y}$ will transform $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ into $\dot{\mathbf{y}} = \mathbf{D}\mathbf{y}$ where $\mathbf{D}$ is a diagonal matrix, or

$$\dot{\mathbf{y}} = \mathbf{U}^{-1}\mathbf{A}\mathbf{U}\mathbf{y} = \mathbf{U}^*\mathbf{A}\mathbf{U}\mathbf{y} = \mathbf{D}\mathbf{y}$$

In fact,

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} \dfrac{j\sqrt{3}}{\sqrt{6}} & 0 & \dfrac{1+j\sqrt{2}}{\sqrt{6}} \\ \dfrac{1-j\sqrt{2}}{\sqrt{10}} & \dfrac{-2}{\sqrt{10}} & \dfrac{j\sqrt{3}}{\sqrt{10}} \\ \dfrac{1-j\sqrt{2}}{\sqrt{15}} & \dfrac{3}{\sqrt{15}} & \dfrac{j\sqrt{3}}{\sqrt{15}} \end{bmatrix} \begin{bmatrix} 1 & 1+j\sqrt{2} & 0 \\ 1-j\sqrt{2} & 2 & j\sqrt{3} \\ 0 & -j\sqrt{3} & 1 \end{bmatrix}$$

$$\cdot \begin{bmatrix} \dfrac{-j\sqrt{3}}{\sqrt{6}} & \dfrac{1+j\sqrt{2}}{\sqrt{10}} & \dfrac{1+j\sqrt{2}}{\sqrt{15}} \\ 0 & \dfrac{-2}{\sqrt{10}} & \dfrac{3}{\sqrt{15}} \\ \dfrac{1-j\sqrt{2}}{\sqrt{6}} & \dfrac{-j\sqrt{3}}{\sqrt{10}} & \dfrac{-j\sqrt{3}}{\sqrt{15}} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

The initial condition is transformed into

$$\mathbf{y}(0) = \mathbf{U}^{-1}\mathbf{x}(0) = \mathbf{U}^*\mathbf{x}(0)$$

With trivial modifications to Theorem 3–5 we can prove that real symmetric matrices are orthogonally similar to diagonal matrices. (The proof is given in Prob. A-3-3.) In the following we prove that if $\mathbf{A}$ is real and is orthogonally similar to a diagonal matrix, then $\mathbf{A}$ is symmetric. In other words if $\mathbf{A}$ is over the field of real numbers, then only real symmetric matrices are orthogonally similar to diagonal matrices.

Theorem 3-6 An $n \times n$ matrix $\mathbf{A}$ over the field of real numbers is orthogonally similar to a diagonal matrix if and only if it is symmetric.

Proof The first part (sufficient condition) of the proof is quite similar to the corresponding part of Theorem 3–5. It is also proved in Prob. A-3-3. Hence, here we shall prove only that if an $n \times n$ real matrix $\mathbf{A}$ is orthogonally similar to a diagonal matrix, then $\mathbf{A}$ is symmetric.

If $\mathbf{A}$ is orthogonally similar to a diagonal matrix, then there exists an orthogonal matrix $\mathbf{T}$ such that

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \mathbf{D}$$

Since $\mathbf{T}$ is orthogonal, $\mathbf{T}^{-1} = \mathbf{T}'$ and we obtain

$$\mathbf{A} = \mathbf{T}\mathbf{D}\mathbf{T}' = \mathbf{T}\mathbf{D}'\mathbf{T}' = \mathbf{A}'$$

Hence $\mathbf{A}$ is symmetric.

Congruent matrices For $n \times n$ matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{P}$, where $\mathbf{P}$ is nonsingular, if the following relationship exists

$$\mathbf{P}'\mathbf{A}\mathbf{P} = \mathbf{B} \quad \text{or} \quad \mathbf{A} = \mathbf{P}\mathbf{B}\mathbf{P}' = (\mathbf{P}')'\mathbf{B}(\mathbf{P}')$$

then $\mathbf{A}$ and $\mathbf{B}$ are said to be *congruent*.

Example 3-2 Obtain a diagonal matrix similar to the following real symmetric matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} 1 & -2 & 2 \\ -2 & 3 & 0 \\ 2 & 0 & 3 \end{bmatrix}$$

Notice that such a real symmetric matrix $\mathbf{A}$ can be diagonalized by the use of the transformation matrix whose columns consist of the normalized eigenvectors of $\mathbf{A}$. These normalized eigenvectors are linearly independent and are orthonormal to each other.

The eigenvalues of $\mathbf{A}$ are obtained as

$$\lambda_1 = -1, \qquad \lambda_2 = 3, \qquad \lambda_3 = 5$$

The eigenvectors corresponding to λ_1, λ_2, and λ_3 are obtained from

$$(\mathbf{A} - \lambda_i\mathbf{I})\mathbf{P}_i = \mathbf{0} \qquad (i = 1, 2, 3)$$

Namely, for λ_1

$$\begin{aligned} 2p_{11} - 2p_{21} + 2p_{31} &= 0 \\ -2p_{11} + 4p_{21} &= 0 \\ 2p_{11} + 4p_{31} &= 0 \end{aligned}$$

from which p_{11}, p_{21}, and p_{31} are obtained as

$$\mathbf{P}_1 = \begin{bmatrix} p_{11} \\ p_{21} \\ p_{31} \end{bmatrix} = \begin{bmatrix} -2a \\ -a \\ a \end{bmatrix}, \qquad a = \text{nonzero constant}$$

Similarly we obtain eigenvectors corresponding to λ_2 and λ_3 as follows:

$$\mathbf{P}_2 = \begin{bmatrix} 0 \\ b \\ b \end{bmatrix}, \qquad b = \text{nonzero constant}$$

$$\mathbf{P}_3 = \begin{bmatrix} c \\ -c \\ c \end{bmatrix}, \qquad c = \text{nonzero constant}$$

Hence

$$\mathbf{P} = [\mathbf{P}_1 \ \mathbf{P}_2 \ \mathbf{P}_3] = \begin{bmatrix} -2a & 0 & c \\ -a & b & -c \\ a & b & c \end{bmatrix}$$

Normalizing each column vector of $\mathbf{P}$ we obtain an orthogonal matrix $\mathbf{T}$ that will transform $\mathbf{A}$ into the diagonal form as follows:

$$\mathbf{T} = \begin{bmatrix} -\frac{2}{\sqrt{6}} & 0 & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \end{bmatrix}$$

In fact,

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \mathbf{T}'\mathbf{A}\mathbf{T}$$

$$= \begin{bmatrix} -\frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix} \begin{bmatrix} 1 & -2 & 2 \\ -2 & 3 & 0 \\ 2 & 0 & 3 \end{bmatrix} \begin{bmatrix} -\frac{2}{\sqrt{6}} & 0 & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \end{bmatrix}$$

$$= \begin{bmatrix} -1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

(For the given $\mathbf{A}$ in this example we need not use the orthogonal matrix $\mathbf{T}$ to transform $\mathbf{A}$ into the diagonal form. The transformation matrix $\mathbf{P}$ defined by Eq. (3–18) will transform $\mathbf{A}$ into the same diagonal form, since the given $\mathbf{A}$ involves only distinct eigenvalues. If a real symmetric matrix $\mathbf{A}$ involves multiple eigenvalues, then we may use Method 2 on page 118 or we may use an orthogonal matrix $\mathbf{T}$ for transforming $\mathbf{A}$ into the diagonal form. See the following example.)

Example 3-3 Show that the following real symmetric matrix $\mathbf{A}$ is orthogonally similar to a diagonal matrix:

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & -1 \\ 1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}$$

The eigenvalues of $\mathbf{A}$ are found to be

$$\lambda_1 = 1, \qquad \lambda_2 = 1, \qquad \lambda_3 = 4$$

Eigenvectors corresponding to $\lambda_1 = \lambda_2 = 1$ are obtained by solving

$$(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{P}_1 = \mathbf{0}$$

or

$$\begin{aligned} p_{11} + p_{21} - p_{31} &= 0 \\ p_{11} + p_{21} - p_{31} &= 0 \\ -p_{11} - p_{21} + p_{31} &= 0 \end{aligned}$$

Notice that these three equations are identical; hence we have three unknowns and one equation. This equation represents the eigenvectors corresponding to $\lambda_1 = \lambda_2 = 1$. We may arbitrarily choose $p_{11} = 0, p_{21} = a$. Then $p_{31} = a$. Thus

$$\mathbf{P}_1 = \begin{bmatrix} p_{11} \\ p_{21} \\ p_{31} \end{bmatrix} = \begin{bmatrix} 0 \\ a \\ a \end{bmatrix}, \qquad a = \text{nonzero constant}$$

The second eigenvector $\mathbf{P}_2$ is obtained so that it is orthogonal to $\mathbf{P}_1$ and satisfies the equation

$$p_{12} + p_{22} - p_{32} = 0$$

From the orthogonality condition

$$(\mathbf{P}_1, \mathbf{P}_2) = 0$$

we obtain

$$p_{22} + p_{32} = 0$$

Hence we obtain $\mathbf{P}_2$ as

$$\mathbf{P}_2 = \begin{bmatrix} p_{12} \\ p_{22} \\ p_{32} \end{bmatrix} = \begin{bmatrix} 2b \\ -b \\ b \end{bmatrix}, \qquad b = \text{nonzero constant}$$

For the eigenvalue $\lambda_3 = 4$, we obtain the eigenvector $\mathbf{P}_3$ as

$$\mathbf{P}_3 = \begin{bmatrix} p_{13} \\ p_{23} \\ p_{33} \end{bmatrix} = \begin{bmatrix} c \\ c \\ -c \end{bmatrix}, \qquad c = \text{nonzero constant}$$

Hence

$$\mathbf{P} = [\mathbf{P}_1 \ \mathbf{P}_2 \ \mathbf{P}_3] = \begin{bmatrix} 0 & 2b & c \\ a & -b & c \\ a & b & -c \end{bmatrix}$$

Normalizing the column vectors of $\mathbf{P}$ we obtain the following orthogonal matrix $\mathbf{T}$:

$$\mathbf{T} = \begin{bmatrix} 0 & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{3}} \end{bmatrix}$$

The orthogonal matrix $\mathbf{T}$ thus obtained will transform $\mathbf{A}$ into the diagonal form. In fact,

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \mathbf{T}'\mathbf{A}\mathbf{T}$$

$$= \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \end{bmatrix} \begin{bmatrix} 2 & 1 & -1 \\ 1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} 0 & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{3}} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

3-4 SYSTEMS WITH MULTIPLE EIGENVECTORS—REDUCTION OF COEFFICIENT MATRICES INTO THE JORDAN CANONICAL FORM (3-I)[1]

Consider the following five matrices:

$$\begin{bmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 \\ 0 & 0 & \lambda_1 & 1 \\ 0 & 0 & 0 & \lambda_1 \end{bmatrix}, \quad \begin{bmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 \\ 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & \lambda_1 \end{bmatrix}, \quad \begin{bmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & \lambda_1 & 1 \\ 0 & 0 & 0 & \lambda_1 \end{bmatrix}$$

$$\begin{bmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & \lambda_1 \end{bmatrix}, \quad \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & \lambda_1 \end{bmatrix}$$

Each of the five preceding matrices has the same characteristic equation $(\lambda - \lambda_1)^4 = 0$. The first one corresponds to the case where there exists only one linearly independent eigenvector, since by denoting the first matrix by $\mathbf{A}$ and solving the following equation:

$$(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{x} = \mathbf{0}$$

we obtain only one eigenvector

$$\mathbf{x} = \begin{bmatrix} a \\ 0 \\ 0 \\ 0 \end{bmatrix}, \qquad a = \text{nonzero constant}$$

The second, third, fourth, and fifth matrices have, respectively, two, two, three, and four linearly independent eigenvectors. (Notice that only the diagonal matrix has four linearly independent eigenvectors.)

If an $n \times n$ matrix $\mathbf{A}$ does not possess n linearly independent eigenvectors, then $\mathbf{A}$ is not similar to a diagonal matrix. It can be proved rigorously that an $n \times n$ matrix $\mathbf{A}$ possessing less than n linearly independent eigenvectors is similar to the Jordan canonical form, where the elements on the main diagonal are the eigenvalues and the elements immediately above the main diagonal are either 1 or 0 and all other elements are zero. We shall not prove this statement in this book,[1] but shall use the result.

In this section we shall focus our attention on the matrix in which the multiple eigenvalue of multiplicity m yields a single eigenvector of multiplicity m. We shall obtain transformation matrices that transform matrices with such multiple eigenvectors into the Jordan canonical form. For example, if $\mathbf{A}$ involves eigenvalues λ_1, λ_1, λ_1, λ_4, and λ_5, we shall find a transformation matrix $\mathbf{P}$ such that

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{bmatrix} \lambda_1 & 1 & 0 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & 0 & \lambda_4 & 0 \\ 0 & 0 & 0 & 0 & \lambda_5 \end{bmatrix}$$

Notice that if we merely follow the method of transformation discussed in Section 3–2 for systems with distinct eigenvalues, then a difficulty arises because the matrix $\mathbf{P}$ given by Eq. (3–18) becomes singular and $\mathbf{P}^{-1}$ does not exist. Therefore $\mathbf{P}$ matrix must be modified.

We shall use a third-order system as an illustrative example and then extend the results obtained to more general cases. Consider a third-order system

[1]See Reference at end of chapter.

[1]Proofs of this statement may be found in a number of books on matrices. See for example (3–2).

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} \tag{3-23}$$

or

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

If the multiple eigenvectors are double eigenvectors, then we may first assume that the corresponding double eigenvalues are slightly different from each other and write eigenvalues of $\mathbf{A}$ as λ_1, $\lambda_1 + \Delta\lambda_1$, λ_3. We then construct a transformation matrix and let $\Delta\lambda_1 \longrightarrow 0$. If we denote a transformation matrix defined by Eq. (3–18) as $\mathbf{P}$, then

$$\mathbf{P} = \begin{bmatrix} A_{111} & A_{211} & A_{311} \\ A_{112} & A_{212} & A_{312} \\ A_{113} & A_{213} & A_{313} \end{bmatrix}$$

By letting $\lambda_2 = \lambda_1 + \Delta\lambda_1$ we obtain

$$\mathbf{P} = \begin{bmatrix} \begin{vmatrix} e-\lambda_1 & f \\ h & i-\lambda_1 \end{vmatrix} & \begin{vmatrix} e-\lambda_1-\Delta\lambda_1 & f \\ h & i-\lambda_1-\Delta\lambda_1 \end{vmatrix} & \begin{vmatrix} e-\lambda_3 & f \\ h & i-\lambda_3 \end{vmatrix} \\ -\begin{vmatrix} d & f \\ g & i-\lambda_1 \end{vmatrix} & -\begin{vmatrix} d & f \\ g & i-\lambda_1-\Delta\lambda_1 \end{vmatrix} & -\begin{vmatrix} d & f \\ g & i-\lambda_3 \end{vmatrix} \\ \begin{vmatrix} d & e-\lambda_1 \\ g & h \end{vmatrix} & \begin{vmatrix} d & e-\lambda_1-\Delta\lambda_1 \\ g & h \end{vmatrix} & \begin{vmatrix} d & e-\lambda_3 \\ g & h \end{vmatrix} \end{bmatrix}$$

By the transformation $\mathbf{x} = \mathbf{P}\mathbf{y}$, we obtain

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_1 + \Delta\lambda_1 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

If we perform the second transformation defined by $\mathbf{y} = \mathbf{Q}\mathbf{z}$, where

$$\mathbf{Q} = \begin{bmatrix} 0 & -\dfrac{1}{\Delta\lambda_1} & 0 \\ 1 & \dfrac{1}{\Delta\lambda_1} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

the system equation becomes

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 + \Delta\lambda_1 & 1 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \tag{3-24}$$

Now let $\Delta\lambda$ approach zero, then Eq. (3–24) becomes:

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 1 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \tag{3-25}$$

Equation (3–25) is the simplest form we can get when the system of Eq. (3–23) involves double eigenvectors. (It is noted that it is impossible to separate the variables z_1 from z_2.)

The transformation matrix $\mathbf{S}$ that transforms Eq. (3–23) into Eq. (3–25) can be obtained as follows:

$$\mathbf{S} = \lim_{\Delta\lambda_1 \to 0} \mathbf{P}\mathbf{Q}$$

$$= \lim_{\Delta\lambda_1 \to 0} \begin{bmatrix} \begin{vmatrix} e-\lambda_1-\Delta\lambda_1 & f \\ h & i-\lambda_1-\Delta\lambda_1 \end{vmatrix} & 2\lambda_1+\Delta\lambda_1-(i+e) & \begin{vmatrix} e-\lambda_3 & f \\ h & i-\lambda_3 \end{vmatrix} \\ -\begin{vmatrix} d & f \\ g & i-\lambda_1-\Delta\lambda_1 \end{vmatrix} & d & -\begin{vmatrix} d & f \\ g & i-\lambda_3 \end{vmatrix} \\ \begin{vmatrix} d & e-\lambda_1-\Delta\lambda_1 \\ g & h \end{vmatrix} & g & \begin{vmatrix} d & e-\lambda_3 \\ g & h \end{vmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} \begin{vmatrix} e-\lambda_1 & f \\ h & i-\lambda_1 \end{vmatrix} & 2\lambda_1-(i+e) & \begin{vmatrix} e-\lambda_3 & f \\ h & i-\lambda_3 \end{vmatrix} \\ -\begin{vmatrix} d & f \\ g & i-\lambda_1 \end{vmatrix} & d & -\begin{vmatrix} d & f \\ g & i-\lambda_3 \end{vmatrix} \\ \begin{vmatrix} d & e-\lambda_1 \\ g & h \end{vmatrix} & g & \begin{vmatrix} d & e-\lambda_3 \\ g & h \end{vmatrix} \end{bmatrix} \tag{3-26}$$

Notice that the elements in the second column are derivatives (with respect to λ_1) of the corresponding elements in the first column. (Note that matrix $\mathbf{S}$ can always be made nonsingular. See Example 3-4 and Prob. A-3-5.)

In the foregoing we have considered the case where the eigenvectors of $\mathbf{A}$ involve double eigenvectors. We shall next consider the case where three eigenvectors of $\mathbf{A}$ are all equal, or $\mathbf{A}$ has only one linearly independent eigenvector. In this case the three eigenvalues are equal, or $\lambda_1 = \lambda_2 = \lambda_3$. In such case we can proceed exactly as we have done for the double eigenvector system. We first let $\lambda_2 = \lambda_1 + \Delta\lambda_1$, $\lambda_3 = \lambda_1 + 2\Delta\lambda_1$ and define $\mathbf{P}$ as follows:

$$\mathbf{P} = \begin{bmatrix} \begin{vmatrix} e-\lambda_1 & f \\ h & i-\lambda_1 \end{vmatrix} & \begin{vmatrix} e-\lambda_1-\Delta\lambda_1 & f \\ h & i-\lambda_1-\Delta\lambda_1 \end{vmatrix} & \begin{vmatrix} e-\lambda_1-2\Delta\lambda_1 & f \\ h & i-\lambda_1-2\Delta\lambda_1 \end{vmatrix} \\ -\begin{vmatrix} d & f \\ g & i-\lambda_1 \end{vmatrix} & -\begin{vmatrix} d & f \\ g & i-\lambda_1-\Delta\lambda_1 \end{vmatrix} & -\begin{vmatrix} d & f \\ g & i-\lambda_1-2\Delta\lambda_1 \end{vmatrix} \\ \begin{vmatrix} d & e-\lambda_1 \\ g & h \end{vmatrix} & \begin{vmatrix} d & e-\lambda_1-\Delta\lambda_1 \\ g & h \end{vmatrix} & \begin{vmatrix} d & e-\lambda_1-2\Delta\lambda_1 \\ g & h \end{vmatrix} \end{bmatrix}$$

Then, the transformation $\mathbf{x} = \mathbf{Py}$ will transform the system equation into

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_1 + \Delta\lambda_1 & 0 \\ 0 & 0 & \lambda_1 + 2\Delta\lambda_1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

If we use the second transformation $\mathbf{y} = \mathbf{Qz}$, where

$$\mathbf{Q} = \begin{bmatrix} 0 & 0 & \dfrac{1}{2\Delta\lambda_1^2} \\ 0 & -\dfrac{1}{\Delta\lambda_1} & -\dfrac{1}{\Delta\lambda_1^2} \\ 1 & \dfrac{1}{\Delta\lambda_1} & \dfrac{1}{2\Delta\lambda_1^2} \end{bmatrix}$$

then we obtain

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 + 2\Delta\lambda_1 & 1 & 0 \\ 0 & \lambda_1 + \Delta\lambda_1 & 1 \\ 0 & 0 & \lambda_1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \tag{3-27}$$

Now let $\Delta\lambda_1$ approach zero, then Eq. (3–27) becomes:

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 1 & 0 \\ 0 & \lambda_1 & 1 \\ 0 & 0 & \lambda_1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \tag{3-28}$$

The transformation matrix $\mathbf{S}$ that transforms Eq. (3–23) into Eq. (3–28), the simplest form for the system with $\lambda_1 = \lambda_2 = \lambda_3$, can be obtained as follows:

$$\mathbf{S} = \lim_{\Delta\lambda_1 \to 0} \mathbf{PQ}$$

$$= \lim_{\Delta\lambda_1 \to 0} \begin{bmatrix} \begin{vmatrix} e-\lambda_1-2\Delta\lambda_1 & f \\ h & i-\lambda_1-2\Delta\lambda_1 \end{vmatrix} & 2\lambda_1+2\Delta\lambda_1-(i+e) & 1 \\ -\begin{vmatrix} d & f \\ g & i-\lambda_1-2\Delta\lambda_1 \end{vmatrix} & d & 0 \\ \begin{vmatrix} d & e-\lambda_1-2\Delta\lambda_1 \\ g & h \end{vmatrix} & g & 0 \end{bmatrix}$$

$$= \begin{bmatrix} \begin{vmatrix} e-\lambda_1 & f \\ h & i-\lambda_1 \end{vmatrix} & 2\lambda_1-(i+e) & 1 \\ -\begin{vmatrix} d & f \\ g & i-\lambda_1 \end{vmatrix} & d & 0 \\ \begin{vmatrix} d & e-\lambda_1 \\ g & h \end{vmatrix} & g & 0 \end{bmatrix} \tag{3-29}$$

Notice that in Eq. (3–29) the first column consists of A_{11j}, the second column is the first derivative (with respect to λ_1) of the first column, and the third column is the second derivative of the first column divided by two.

The preceding analysis can be extended to general cases. If the characteristic equation involves, for example, seven eigenvalues ($\lambda_1, \lambda_1, \lambda_1, \lambda_1, \lambda_5, \lambda_5, \lambda_7$) but has only three linearly independent eigenvectors, then the transformation matrix $\mathbf{S}$ that transforms the original coefficient matrix $\mathbf{A}$ into the Jordan canonical form can be given as follows:

$$\mathbf{S} = \begin{bmatrix} A_{111} & A'_{111} & \frac{1}{2!}A''_{111} & \frac{1}{3!}A'''_{111} & A_{511} & A'_{511} & A_{711} \\ A_{112} & A'_{112} & \frac{1}{2!}A''_{112} & \frac{1}{3!}A'''_{112} & A_{512} & A'_{512} & A_{712} \\ A_{113} & A'_{113} & \frac{1}{2!}A''_{113} & \frac{1}{3!}A'''_{113} & A_{513} & A'_{513} & A_{713} \\ A_{114} & A'_{114} & \frac{1}{2!}A''_{114} & \frac{1}{3!}A'''_{114} & A_{514} & A'_{514} & A_{714} \\ A_{115} & A'_{115} & \frac{1}{2!}A''_{115} & \frac{1}{3!}A'''_{115} & A_{515} & A'_{515} & A_{715} \\ A_{116} & A'_{116} & \frac{1}{2!}A''_{116} & \frac{1}{3!}A'''_{116} & A_{516} & A'_{516} & A_{716} \\ A_{117} & A'_{117} & \frac{1}{2!}A''_{117} & \frac{1}{3!}A'''_{117} & A_{517} & A'_{517} & A_{717} \end{bmatrix} \tag{3-30}$$

where primes denote derivatives with respect to λ_i, for example,

$$A'''_{111} = \frac{d^3}{d\lambda_1^3}A_{111}(\lambda_1)$$

$$A''_{113} = \frac{d^2}{d\lambda_1^2}A_{113}(\lambda_1)$$

$$A'_{515} = \frac{d}{d\lambda_5}A_{515}(\lambda_5)$$

If the transformation $\mathbf{x} = \mathbf{Sz}$ is carried out, then the transformed equation will have the following Jordan canonical form:

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \\ \dot{z}_4 \\ \dot{z}_5 \\ \dot{z}_6 \\ \dot{z}_7 \end{bmatrix} = \left[\begin{array}{cccc:cc:c} \lambda_1 & 1 & 0 & 0 & & & 0 \\ 0 & \lambda_1 & 1 & 0 & & & \\ 0 & 0 & \lambda_1 & 1 & & & \\ 0 & 0 & 0 & \lambda_1 & & & \\ \hdashline & & & & \lambda_5 & 1 & \\ & & & & 0 & \lambda_5 & \\ \hdashline 0 & & & & & & \lambda_7 \end{array}\right] \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ z_6 \\ z_7 \end{bmatrix}$$

Similarly, for the system having m identical eigenvalues $\lambda_1 = \lambda_2 = \cdots = \lambda_m$, for which there exists only one linearly independent eigenvector, the first column of $\mathbf{S}$ consists of A_{11j}, the second column consists of the first derivative (with respect to λ_1) of the first column, and the ith column consists of $(i-1)$th derivative of the first column divided by $(i-1)!$, $(i = 2, 3, \ldots, m)$. If, in numerical computation, all the elements of the first column turn out to be zero, then we substitute A_{1kj} $(k \neq 1; j = 1, 2, \ldots, n)$ for the first column, and the ith column is then the $(i-1)$th derivative of the first column divided by $(i-1)!$, $(i = 2, 3, \ldots, m)$. Similarly, if all the elements A_{p1j} $(j = 1, 2, \ldots, n)$ of the pth column (corresponding to the eigenvalue λ_p) turn out to be zero, then we substitute A_{pkj} $(k \neq 1)$ for the pth column.

Note that such a transformation matrix derived in this section is not unique for a given coefficient matrix $\mathbf{A}$. Suppose that an $n \times n$ matrix $\mathbf{S}$ is such that $\mathbf{S}^{-1}\mathbf{AS} = \mathbf{J}$, where $\mathbf{J}$ is a matrix in the Jordan canonical form. If a nonsingular matrix $\mathbf{T}$ is equal to $\mathbf{SV}$ where $\mathbf{V}$ is nonsingular and satisfies $\mathbf{JV} = \mathbf{VJ}$; then $\mathbf{T}^{-1}\mathbf{AT} = \mathbf{J}$. Clearly, since $\mathbf{T}$ equals $\mathbf{SV}$, we have

$$\mathbf{TV}^{-1} = \mathbf{S} \qquad \text{and} \qquad \mathbf{VT}^{-1}\mathbf{ATV}^{-1} = \mathbf{J}$$

or

$$\mathbf{T}^{-1}\mathbf{AT} = \mathbf{V}^{-1}\mathbf{JV} = \mathbf{V}^{-1}\mathbf{VJ} = \mathbf{J}$$

We thus see that matrix $\mathbf{S}$ is not unique. Consider as an example the following $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 3 \\ 4 & 1 & 2 \\ -\frac{24}{13} & \frac{10}{13} & 2 \end{bmatrix}$$

The characteristic polynominal is

$$|\mathbf{A} - \lambda\mathbf{I}| = -(\lambda - 2)(\lambda - 2)(\lambda - 1)$$

A transformation matrix $\mathbf{S}$ can be obtained from Eq. (3–26) as follows:

$$\mathbf{S} = \begin{bmatrix} -\frac{20}{13} & 1 & -\frac{20}{13} \\ -\frac{48}{13} & 4 & -\frac{100}{13} \\ \frac{16}{13} & -\frac{24}{13} & \frac{40}{13} \end{bmatrix}$$

This matrix $\mathbf{S}$ will transform $\mathbf{A}$ into the Jordan canonical form, or $\mathbf{S}^{-1}\mathbf{AS} = \mathbf{J}$, where

$$\mathbf{J} = \begin{bmatrix} \lambda_1 & 1 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \qquad (\lambda_1 = 2, \qquad \lambda_3 = 1)$$

The matrix $\mathbf{V}$ which satisfies $\mathbf{JV} = \mathbf{VJ}$ will have the following form:

$$\mathbf{V} = \begin{bmatrix} v_1 & v_2 & 0 \\ 0 & v_1 & 0 \\ 0 & 0 & v_3 \end{bmatrix}$$

Here v_1, v_2, and v_3 are arbitrary constants satisfying $v_1 v_3 \neq 0$. If we choose, for example,

$$v_1 = -\tfrac{13}{48}, \qquad v_2 = -\tfrac{3}{2}\cdot\tfrac{13}{48}, \qquad v_3 = \tfrac{13}{40}$$

then the following matrix $\mathbf{T}$

$$\mathbf{T} = \mathbf{SV} = \begin{bmatrix} \frac{5}{12} & \frac{17}{48} & -\frac{1}{2} \\ 1 & \frac{5}{12} & -\frac{5}{2} \\ -\frac{1}{3} & 0 & 1 \end{bmatrix}$$

will transform $\mathbf{A}$ into the same Jordan canonical form, or

$$\mathbf{T}^{-1}\mathbf{AT} = \mathbf{J}$$

Example 3-4 Consider the following matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} 5 & 4 & 0 \\ 0 & 1 & 0 \\ -4 & 4 & 1 \end{bmatrix}$$

The eigenvalues of $\mathbf{A}$ are obtained as

$$\lambda_1 = 1, \qquad \lambda_2 = 1, \qquad \lambda_3 = 5$$

For the eigenvalue $\lambda_3 = 5$, the corresponding eigenvector is obtained by solving

$$(\mathbf{A} - 5\mathbf{I})\mathbf{x}_3 = \mathbf{0}$$

for $\mathbf{x}_3$, or solving

$$\begin{aligned} 4x_{23} &= 0 \\ -4x_{23} &= 0 \\ -4x_{13} + 4x_{23} - 4x_{33} &= 0 \end{aligned}$$

for x_{13}, x_{23}, and x_{33}. The result is

$$\begin{bmatrix} x_{13} \\ x_{23} \\ x_{33} \end{bmatrix} = \begin{bmatrix} -a \\ 0 \\ a \end{bmatrix}, \qquad a = \text{nonzero constant}$$

For the multiple eigenvalue $\lambda_1 = \lambda_2 = 1$, the corresponding eigenvectors may be obtained by solving

$$(\mathbf{A} - \mathbf{I})\mathbf{x}_1 = \mathbf{0}$$

or

$$\begin{aligned} 4x_{11} + 4x_{21} &= 0 \\ 0x_{11} + 0x_{21} + 0x_{31} &= 0 \\ -4x_{11} + 4x_{21} &= 0 \end{aligned}$$

from which we obtain only one linearly independent eigenvector

$$\begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ b \end{bmatrix}, \qquad b = \text{nonzero constant}$$

This means that the given matrix **A** possesses only two linearly independent eigenvectors. Hence **A** can not be diagonalized. The simplest form we can obtain is the Jordan canonical form:

$$\mathbf{J} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

Let us obtain by use of Eq. (3–26) a transformation matrix **S** such that

$$\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{J}$$

Notice that the first column of **S** given by Eq. (3–26) becomes zero, or

$$\begin{bmatrix} \begin{vmatrix} e-\lambda_1 & f \\ h & i-\lambda_1 \end{vmatrix} \\ -\begin{vmatrix} d & f \\ g & i-\lambda_1 \end{vmatrix} \\ \begin{vmatrix} d & e-\lambda_1 \\ g & h \end{vmatrix} \end{bmatrix} = \begin{bmatrix} \begin{vmatrix} 0 & 0 \\ 4 & 0 \end{vmatrix} \\ -\begin{vmatrix} 0 & 0 \\ -4 & 0 \end{vmatrix} \\ \begin{vmatrix} 0 & 0 \\ -4 & 4 \end{vmatrix} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Hence we modify Eq. (3–26) as follows:

$$\mathbf{S} = \begin{bmatrix} A_{121}(\lambda_1) & \dfrac{d}{d\lambda_1}A_{121}(\lambda_1) & A_{311}(\lambda_3) \\ A_{122}(\lambda_1) & \dfrac{d}{d\lambda_1}A_{122}(\lambda_1) & A_{312}(\lambda_3) \\ A_{123}(\lambda_1) & \dfrac{d}{d\lambda_1}A_{123}(\lambda_1) & A_{313}(\lambda_3) \end{bmatrix}$$

$$= \begin{bmatrix} -\begin{vmatrix} b & c \\ h & i-\lambda_1 \end{vmatrix} & b & \begin{vmatrix} e-\lambda_3 & f \\ h & i-\lambda_3 \end{vmatrix} \\ \begin{vmatrix} a-\lambda_1 & c \\ g & i-\lambda_1 \end{vmatrix} & 2\lambda_1-(a+i) & -\begin{vmatrix} d & f \\ g & i-\lambda_3 \end{vmatrix} \\ -\begin{vmatrix} a-\lambda_1 & b \\ g & h \end{vmatrix} & h & \begin{vmatrix} d & e-\lambda_3 \\ g & h \end{vmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} -\begin{vmatrix} 4 & 0 \\ 4 & 0 \end{vmatrix} & 4 & \begin{vmatrix} -4 & 0 \\ 4 & -4 \end{vmatrix} \\ \begin{vmatrix} 4 & 0 \\ -4 & 0 \end{vmatrix} & -4 & -\begin{vmatrix} 0 & 0 \\ -4 & -4 \end{vmatrix} \\ -\begin{vmatrix} 4 & 4 \\ -4 & 4 \end{vmatrix} & 4 & \begin{vmatrix} 0 & -4 \\ -4 & 4 \end{vmatrix} \end{bmatrix} = \begin{bmatrix} 0 & 4 & 16 \\ 0 & -4 & 0 \\ -32 & 4 & -16 \end{bmatrix}$$

$\mathbf{S}^{-1}$ is obtained as

$$\mathbf{S}^{-1} = \begin{bmatrix} -\frac{1}{3\cdot 2} & -\frac{1}{1\cdot 6} & -\frac{1}{3\cdot 2} \\ 0 & -\frac{1}{4} & 0 \\ \frac{1}{1\cdot 6} & \frac{1}{1\cdot 6} & 0 \end{bmatrix}$$

and

$$\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \begin{bmatrix} -\frac{1}{3\cdot 2} & -\frac{1}{1\cdot 6} & -\frac{1}{3\cdot 2} \\ 0 & -\frac{1}{4} & 0 \\ \frac{1}{1\cdot 6} & \frac{1}{1\cdot 6} & 0 \end{bmatrix} \begin{bmatrix} 5 & 4 & 0 \\ 0 & 1 & 0 \\ -4 & 4 & 1 \end{bmatrix} \begin{bmatrix} 0 & 4 & 16 \\ 0 & -4 & 0 \\ -32 & 4 & -16 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

3-5 MACHINE COMPUTATION OF TRANSFORMATION MATRICES[1]

The hand computation of the transformation matrix **P** or **S** discussed in Sections 3–2, 3–3, and 3–4 for an $n \times n$ matrix **A** becomes laborious if $n = 4$ or 5 and becomes impractical if $n > 5$. For $n > 5$ machine computation may become a necessity.

This section presents the principle of machine computation of the transformation matrix. The process of computing the transformation matrix starts with computation of the eigenvalues of **A**. This can be done by a standard technique (*3–6*). Hence we assume that the eigenvalues of **A** are already found.

We shall first consider the case where the Jordan canonical form of **A** is diagonal. Let us denote the transformation matrix by **P**, or

$$\mathbf{P} = [\mathbf{P}_1 \,|\, \mathbf{P}_2 \,|\, \cdots \,|\, \mathbf{P}_n]$$

where $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_n$ are as yet unknown eigenvectors. To determine eigenvectors $\mathbf{P}_i$, we successively solve the equation:

$$(\mathbf{A} - \lambda_i\mathbf{I})\mathbf{P}_i = \mathbf{0}$$

Suppose that the eigenvalues of **A** are

$$\lambda_1, \lambda_1, \ldots, \lambda_1, \lambda_{m+1}, \lambda_{m+2}, \ldots, \lambda_n$$

where $\lambda_1, \lambda_{m+1}, \lambda_{m+2}, \ldots, \lambda_n$ are different from each other. (Extension to other cases will become obvious.) If the m eigenvectors corresponding to m equal eigenvalues λ_1 are linearly independent, then the Jordan canonical form of **A** is diagonal. The eigenvectors $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_m$ are found by performing elementary operations on $\mathbf{A} - \lambda_1\mathbf{I}$ until the last m rows of the matrix are identically zero. This is possible since the rank of $\mathbf{A} - \lambda_1\mathbf{I}$ is $n - m$. We shall define the resulting matrix with m rows consisting of identically zero elements as $\mathbf{R}_1$. Then $\mathbf{R}_1$ in a reduced form can be written as

[1]This section closely follows Reference (*3–5*).

$$\mathbf{R}_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 & r_{11} & \cdots & r_{1m} \\ 0 & 1 & \cdots & 0 & r_{21} & \cdots & r_{2m} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 & r_{n-m,1} & \cdots & r_{n-m,m} \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix}$$

The m eigenvectors corresponding to λ_1 are

$$\mathbf{P}_1 = \begin{bmatrix} r_{11} \\ r_{21} \\ \vdots \\ r_{n-m,1} \\ -1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \ldots, \quad \mathbf{P}_m = \begin{bmatrix} r_{1m} \\ r_{2m} \\ \vdots \\ r_{n-m,m} \\ 0 \\ \vdots \\ 0 \\ -1 \end{bmatrix}$$

as may be seen by performing the multiplication $\mathbf{R}_1\mathbf{P}_i$ $(i = 1, 2, \ldots, m)$ and noting that $\mathbf{R}_1\mathbf{P}_i = \mathbf{0}$.

The eigenvectors corresponding to $\lambda_{m+1}, \ldots, \lambda_n$ are found by performing elementary operations on $\mathbf{A} - \lambda_i\mathbf{I}$ $(i = m + 1, \ldots, n)$ until the last row of the matrix is identically zero. This is always possible because $\mathbf{A} - \lambda\mathbf{I}$ is singular for $\lambda = \lambda_i$. If the matrix is reduced to the form

$$\mathbf{R}_i = \begin{bmatrix} 1 & 0 & \cdots & 0 & r_1 \\ 0 & 1 & \cdots & 0 & r_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & r_{n-1} \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

then clearly an eigenvector for λ_i is

$$\mathbf{P}_i = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_{n-1} \\ -1 \end{bmatrix}$$

The transformation matrix $\mathbf{P}$ can thus be determined.

Next, consider the case where the Jordan canonical form of $\mathbf{A}$ is not diagonal. This occurs when the eigenvectors for the multiple eigenvalues of $\mathbf{A}$ are not linearly independent. Suppose that $\mathbf{A}$ involves m equal eigenvalues λ_1 and other eigenvalues are all distinct, or the eigenvalues of $\mathbf{A}$ are

$$\lambda_1, \lambda_1, \ldots, \lambda_1, \lambda_{m+1}, \lambda_{m+2}, \ldots, \lambda_n$$

Suppose that there is only one linearly independent eigenvector for λ_1. [Note that there are other cases where $q(1 < q < m)$ linearly independent eigenvectors exist for λ_1. Analyses of such cases involve no particular difficulty.] Once the eigenvector $\mathbf{S}_1$ is found, the other vectors $\mathbf{S}_2, \mathbf{S}_3, \ldots, \mathbf{S}_m$ may be easily computed from the following relationships:

$$(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{S}_k = \mathbf{S}_{k-1} \qquad (k = 2, 3, \ldots, m)$$

Note that

$$(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{S}_1 = \mathbf{0}$$

because $\mathbf{S}_1$ is an eigenvector for λ_1. From the relationships just given we have

$$(\mathbf{A} - \lambda_1\mathbf{I})^{m-1}\mathbf{S}_m = \mathbf{S}_1 \tag{3-31}$$

and

$$(\mathbf{A} - \lambda_1\mathbf{I})^m\mathbf{S}_m = \mathbf{0} \tag{3-32}$$

Applying the same row reduction process as before to $(\mathbf{A} - \lambda_1\mathbf{I})^m$ we find m linearly independent vectors satisfying Eq. (3–32). These are tested to find one that satisfies Eq. (3–31). $\mathbf{S}_1$ is not known at this point, but any vector on the left side that gives a product not identically zero qualifies as $\mathbf{S}_m$, and the nonzero product is an eigenvector.

We are now ready to construct the transformation matrix. From

$$\mathbf{AS}_i = \lambda_i\mathbf{S}_i \qquad (i = 1, m + 1, m + 2, \ldots, n)$$

$$\mathbf{AS}_k = \mathbf{S}_{k-1} + \lambda_1\mathbf{S}_k \qquad (k = 2, 3, \ldots, m)$$

we see that a matrix consisting of n column vectors $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_n$ satisfies the matrix equation

$$\mathbf{AS} = \mathbf{SJ}$$

where $\mathbf{J}$ is the Jordan canonical form. Since the columns of $\mathbf{S}$ are linearly independent, its inverse exists and it follows that

$$\mathbf{J} = \mathbf{S}^{-1}\mathbf{AS}$$

Note that complex conjugate eigenvalues could be handled in exactly the same way as real eigenvalues by putting complex conjugate columns in the transformation matrix. This, however, wastes computer storage space and calculation time because each of the array must be doubled in size to hold the imaginary components. Rather than doing this it may be advantageous to modify the Jordan canonical form. The modification is based on the fact that one of a pair of complex conjugate vectors contains all of the essential information of the pair.

Let us consider the case where the Jordan canonical form is diagonal.

Other cases can be handled the same way. Notice that the diagonal matrix

$$\mathbf{J} = \begin{bmatrix} \lambda_1 & & & & 0 \\ & \ddots & & & \\ & & \sigma + j\omega & & \\ & & & \sigma - j\omega & \\ & & & & \ddots \\ 0 & & & & \lambda_n \end{bmatrix}$$

can be transformed into the following form

$$\hat{\mathbf{J}} = \begin{bmatrix} \lambda_1 & & & & 0 \\ & \ddots & & & \\ & & \sigma & \omega & \\ & & -\omega & \sigma & \\ & & & & \ddots \\ 0 & & & & \lambda_n \end{bmatrix}$$

by means of the transformation matrix defined by

$$\mathbf{K} = \begin{bmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & \frac{1}{2} & -\frac{j}{2} & \\ & & \frac{1}{2} & \frac{j}{2} & \\ & & & & \ddots \\ 0 & & & & 1 \end{bmatrix}$$

Namely the modified Jordan canonical form $\hat{\mathbf{J}}$ is given as

$$\hat{\mathbf{J}} = \mathbf{K}^{-1}\mathbf{J}\mathbf{K} = \mathbf{K}^{-1}\mathbf{P}^{-1}\mathbf{A}\mathbf{P}\mathbf{K}$$

Not only does $\hat{\mathbf{J}}$ have only real elements but, more significantly, $\mathbf{K}^{-1}\mathbf{P}^{-1}$ and $\mathbf{PK}$ have only real elements. It may be seen that the effect of postmultiplying $\mathbf{P}$ by $\mathbf{K}$ is to set one column equal to the real part, and the next to the imaginary part of the complex eigenvector. Both vectors are entirely real. All other columns of $\mathbf{P}$ are merely copied. $\mathbf{K}^{-1}$ has a similar effect upon the rows of $\mathbf{P}^{-1}$. Setting

$$\hat{\mathbf{P}} = \mathbf{PK}$$

$$\hat{\mathbf{P}}^{-1} = \mathbf{K}^{-1}\mathbf{P}^{-1}$$

the transformation is given the standard form of

$$\hat{\mathbf{J}} = \hat{\mathbf{P}}^{-1}\mathbf{A}\hat{\mathbf{P}}$$

In use we simply enter the real and imaginary parts of the appropriate eigenvector as two adjacent columns of $\mathbf{P}$, and proceed to use it as though it were a true transformation matrix.

3-6 SECOND-ORDER SYSTEMS

Consider the system defined by the following two simultaneous first-order differential equations:

$$\dot{x}_1 = a_{11}x_1 + a_{12}x_2$$

$$\dot{x}_2 = a_{21}x_1 + a_{22}x_2$$

where a_{11}, a_{12}, a_{21}, and a_{22} are constants, or

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{3-33}$$

We shall denote two eigenvalues as λ_1 and λ_2. Since such second-order systems occur very frequently in engineering fields, we shall give special attention to these systems in this section.

According to the coefficient matrix we can divide Eq. (3-33) into the following three cases:

Case 1: The system involves two distinct eigenvalues The following transformation:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a_{22} - \lambda_1 & a_{22} - \lambda_2 \\ -a_{21} & -a_{21} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

or

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \lambda_2 - a_{11} & \lambda_1 - a_{11} \\ -a_{21} & -a_{21} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

will transform Eq. (3-33) into the diagonal form:

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

The initial condition terms are transformed into

$$\begin{bmatrix} y_1(0) \\ y_2(0) \end{bmatrix} = \frac{1}{a_{21}(\lambda_1 - \lambda_2)} \begin{bmatrix} -a_{21} & a_{11} - \lambda_1 \\ a_{21} & \lambda_2 - a_{11} \end{bmatrix} \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix}$$

Case 2: The system involves complex conjugate eigenvalues The following transformation

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}(\lambda_1 + \bar{\lambda}_1 - 2a_{11}) & \frac{j}{2}(\lambda_1 - \bar{\lambda}_1) \\ -a_{21} & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

will transform Eq. (3–33) into the following form:

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

where σ and ω are defined by

$$\lambda_1 = \sigma + j\omega$$
$$\bar{\lambda}_1 = \sigma - j\omega$$

The initial condition terms are transformed into

$$\begin{bmatrix} y_1(0) \\ y_2(0) \end{bmatrix} = \frac{-1}{a_{21}\omega} \begin{bmatrix} 0 & -\frac{j}{2}(\lambda_1 - \bar{\lambda}_1) \\ a_{21} & \frac{1}{2}(\lambda_1 + \bar{\lambda}_1 - 2a_{11}) \end{bmatrix} \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix}$$

Case 3: The system involves two equal eigenvalues For this case, the following two possibilities exist:

(A) If $a_{12} = a_{21} = 0$, $a_{11} = a_{22}$, then the original equation is already in the diagonal form, or

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ 0 & a_{11} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

(B) Except in the special case stated in (A), the following transformation

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a_{11} - \lambda_1 & 1 \\ a_{21} & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

or

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 - a_{22} & 1 \\ a_{21} & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

will transform Eq. (3–33) into the Jordan canonical form:

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 1 \\ 0 & \lambda_1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

The initial condition terms are transformed into

$$\begin{bmatrix} y_1(0) \\ y_2(0) \end{bmatrix} = \frac{1}{a_{21}} \begin{bmatrix} 0 & 1 \\ a_{21} & \lambda_1 - a_{11} \end{bmatrix} \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix}$$

3-7 SYSTEMS DESCRIBED BY SCALAR DIFFERENTIAL EQUATIONS (3–1, 3–3)[1]

Because of frequent occurrence of nth-order scalar differential equations or their equivalent vector matrix differential equations, we shall in this section discuss simplification of state space representation of nth-order scalar differential equations in detail. Consider the following equation:

$$\overset{(n)}{x} + a_1 \overset{(n-1)}{x} + \cdots + a_{n-1}\dot{x} + a_n x = 0 \tag{3–34}$$

where the a_i are constants. If we define

$$x_1 = x$$
$$x_2 = \dot{x}$$
$$\cdots$$
$$x_n = \overset{(n-1)}{x}$$

then Eq. (3–34) can be written as follows:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} \tag{3–35}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix}$$

and

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_2 & -a_1 \end{bmatrix} \tag{3–36}$$

Notice that the eigenvectors of $\mathbf{A}$ defined by Eq. (3–36) become distinct if and only if the eigenvalues are distinct. Stated differently, eigenvectors of $\mathbf{A}$ corresponding to mutliple eigenvalues become multiple. This may be seen easily if we realize that the eigenvector corresponding to an eigenvalue λ_1 is determined from

$$(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{P}_1 = \mathbf{0}$$

or

$$-\lambda_1 p_{11} + p_{21} = 0$$
$$-\lambda_1 p_{21} + p_{31} = 0$$
$$\cdots$$
$$-\lambda_1 p_{n-1,1} + p_{n1} = 0$$
$$-a_n p_{11} - a_{n-1} p_{21} - \cdots - a_1 p_{n1} = 0$$

[1]Figures in parentheses relate to items in list of references at the end of the chapter.

Aside from a constant factor there is only one set of values $p_{11}, p_{21}, \ldots, p_{n1}$, not all of which are equal to zero, that satisfies this set of n equations. This is true whether λ_1 is a distinct eigenvalue or a multiple eigenvalue. Hence the eigenvector corresponding to the mutliple eigenvalue of multiplicity m is also of multiplicity m. The coefficient matrix **A** of the system described by Eq. (3–35) where **A** is given by Eq. (3–36) can be diagonalized if and only if A has distinct eigenvalues.

Nth-order scalar differential equation with distinct eigenvalues Consider Eq. (3–35). We assume that the eigenvalues of **A** are distinct. The transformation matrix **P** given by Eq. (3–18), rewritten thus

$$\mathbf{P} = \begin{bmatrix} A_{111} & A_{211} & \cdots & A_{n11} \\ A_{112} & A_{212} & \cdots & A_{n12} \\ \vdots & \vdots & & \vdots \\ A_{11n} & A_{21n} & \cdots & A_{n1n} \end{bmatrix}$$

can be simplified, since elements A_{i1j} can be obtained in terms of λ_i as follows (Refer to Prob. A-3-6):

$$\begin{aligned} A_{i11} &= \frac{(-)^n a_n}{\lambda_i} \\ A_{i12} &= (-)^n a_n \\ A_{i13} &= (-)^n a_n \lambda_i \qquad (i = 1, 2, \ldots, n) \\ &\cdots \\ A_{i1n} &= (-)^n a_n \lambda_i^{n-2} \end{aligned}$$

Hence the transformation matrix **P** is simplified into

$$\mathbf{P} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_n \\ \lambda_1^2 & \lambda_2^2 & \cdots & \lambda_n^2 \\ \vdots & \vdots & & \vdots \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \cdots & \lambda_n^{n-1} \end{bmatrix} \tag{3–37}$$

After the transformation $\mathbf{x} = \mathbf{P}\mathbf{y}$ is made the coefficient matrix becomes a diagonal matrix and the equation can be written as

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \vdots \\ \dot{y}_n \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

The initial condition is given by

$$\mathbf{y}(0) = \mathbf{P}^{-1}\mathbf{x}(0)$$

It is noted that if some of the λ_i are complex conjugates, for example, $\lambda_2 = \bar{\lambda}_1$ and $\lambda_4 = \bar{\lambda}_3$, then the diagonal elements of the coefficient matrix involve complex elements. In such a case by another transformation the coefficient matrix can be transformed into the following form:

$$\left[\begin{array}{cc:cc} \sigma_1 & \omega_1 & & 0 \\ -\omega_1 & \sigma_1 & & \\ \hdashline & & \sigma_2 & \omega_2 \\ 0 & & -\omega_2 & \sigma_2 \end{array}\right]$$

Nth-order scalar differential equation with multiple eigenvalues We shall next consider Eq. (3–35) with multiple eigenvalues. The transformation matrix which transforms the coefficient matrix into the Jordan canonical form can be obtained by the same technique as employed in Section 3–4. (Refer to Prob. A-3-7.)

Suppose that the matrix **A** in Eq. (3–35) is a 7×7 matrix and the eigenvalues of **A** are $\lambda_1, \lambda_1, \lambda_1, \lambda_1, \lambda_5, \lambda_5, \lambda_7$. Then the transformation matrix is given by Eq. (3–30) in which $A_{111}, A_{112}, \ldots, A_{117}$ are given by

$$\begin{aligned} A_{111} &= 1 \\ A_{112} &= \lambda_1 \\ A_{113} &= \lambda_1^2 \\ A_{114} &= \lambda_1^3 \\ A_{115} &= \lambda_1^4 \\ A_{116} &= \lambda_1^5 \\ A_{117} &= \lambda_1^6 \end{aligned}$$

In other words, the transformation matrix **S** in this case is

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ \lambda_1 & 1 & 0 & 0 & \lambda_5 & 1 & \lambda_6 \\ \lambda_1^2 & 2\lambda_1 & 1 & 0 & \lambda_5^2 & 2\lambda_5 & \lambda_6^2 \\ \lambda_1^3 & 3\lambda_1^2 & 3\lambda_1 & \lambda_1 & \lambda_5^3 & 3\lambda_5^2 & \lambda_6^3 \\ \lambda_1^4 & 4\lambda_1^3 & 6\lambda_1^2 & 4\lambda_1 & \lambda_5^4 & 4\lambda_5^3 & \lambda_6^4 \\ \lambda_1^5 & 5\lambda_1^4 & 10\lambda_1^3 & 10\lambda_1^2 & \lambda_5^5 & 5\lambda_5^4 & \lambda_6^5 \\ \lambda_1^6 & 6\lambda_1^5 & 15\lambda_1^4 & 20\lambda_1^3 & \lambda_5^6 & 6\lambda_5^5 & \lambda_6^6 \end{bmatrix}$$

Notice that the second column is the first derivative (with respect to λ_1) of the first column, the third column is the second derivative of the first column

divided by 2!, the fourth column is the third derivative of the first column divided by 3!; and the sixth column is the first derivative (with respect to λ_5) of the fifth column.

As an example, consider the following system:

$$\dddot{x} + 4\ddot{x} + 5\dot{x} + 2x = 0$$

The eigenvalues of the characteristic equation are -1, -1, and -2. The transformation

$$\mathbf{x} = \mathbf{S}\mathbf{y}$$

where

$$\mathbf{x} = \begin{bmatrix} x \\ \dot{x} \\ \ddot{x} \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 1 \\ \lambda_1 & 1 & \lambda_2 \\ \lambda_1^2 & 2\lambda_1 & \lambda_2^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & -2 \\ 1 & -2 & 4 \end{bmatrix}$$

will transform the original equation into

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

As an another example, consider the following system:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -3 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

The eigenvalues of the coefficient matrix are $\lambda_1 = \lambda_2 = \lambda_3 = 1$. Hence, the transformation:

$$\mathbf{x} = \mathbf{S}\mathbf{y}$$

where

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_1 & 1 & 0 \\ \lambda_1^2 & 2\lambda_1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

will transform the given equation into

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

3-8 MOTIONS OF FREE DYNAMIC SYSTEMS

The solution of the following state space equation:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}, \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{3-38}$$

can be obtained easily through the transformation of $\mathbf{A}$ into the Jordan canonical form. We shall first consider the case where $\mathbf{A}$ has distinct eigenvectors. Since it has been shown that by means of a suitable transformation $\mathbf{x} = \mathbf{P}\mathbf{y}$ we can transform Eq. (3–38) into

$$\dot{\mathbf{y}} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}\mathbf{y} = \mathbf{D}\mathbf{y} \tag{3-39}$$

we shall treat Eq. (3–39) instead of Eq. (3–38).

Consider the following system:

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \vdots \\ \dot{y}_n \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

The solution of this state space equation can be immediately obtained as

$$\begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_n(t) \end{bmatrix} = \begin{bmatrix} e^{\lambda_1 t} y_1(0) \\ e^{\lambda_2 t} y_2(0) \\ \vdots \\ e^{\lambda_n t} y_n(0) \end{bmatrix}$$

or

$$\mathbf{y}(t) = \mathbf{Q}(t)\mathbf{y}(0)$$

where

$$\mathbf{Q}(t) = \begin{bmatrix} e^{\lambda_1 t} & & & 0 \\ & e^{\lambda_2 t} & & \\ & & \ddots & \\ 0 & & & e^{\lambda_n t} \end{bmatrix}$$

Since $\mathbf{y} = \mathbf{P}^{-1}\mathbf{x}$ we obtain

$$\mathbf{y}(0) = \mathbf{P}^{-1}\mathbf{x}(0)$$

Thus, if the solution $\mathbf{x}(t)$ of Eq. (3–38) is desired, it is given by

$$\mathbf{x}(t) = \mathbf{P}\mathbf{y}(t) = \mathbf{P}\mathbf{Q}(t)\mathbf{P}^{-1}\mathbf{x}(0)$$

Example 3-5 Obtain the solution of the following state space equation:

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix}=\begin{bmatrix}-1 & 0 & 0\\ 0 & -2 & 0\\ 0 & 0 & -3\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}$$

The solution is obtained as

$$\begin{bmatrix}x_1(t)\\ x_2(t)\\ x_3(t)\end{bmatrix}=\begin{bmatrix}e^{-t}\,x_1(0)\\ e^{-2t}x_2(0)\\ e^{-3t}x_3(0)\end{bmatrix}$$

or

$$\mathbf{x}(t)=\mathbf{Q}(t)\mathbf{x}(0)$$

where

$$\mathbf{Q}(t)=\begin{bmatrix}e^{-t} & 0 & 0\\ 0 & e^{-2t} & 0\\ 0 & 0 & e^{-3t}\end{bmatrix}$$

Example 3-6 Obtain the solution of the following state space equation:

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix}=\begin{bmatrix}\sigma+j\omega & 0 & 0\\ 0 & \sigma-j\omega & 0\\ 0 & 0 & \lambda_3\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}$$

The solution is obtained as

$$\begin{bmatrix}x_1(t)\\ x_2(t)\\ x_3(t)\end{bmatrix}=\begin{bmatrix}e^{(\sigma+j\omega)t}x_1(0)\\ e^{(\sigma-j\omega)t}x_2(0)\\ e^{\lambda_3 t}\,x_3(0)\end{bmatrix}=\begin{bmatrix}e^{\sigma t}(\cos\omega t+j\sin\omega t)x_1(0)\\ e^{\sigma t}(\cos\omega t-j\sin\bar{\omega}t)x_2(0)\\ e^{\lambda_3 t}x_3(0)\end{bmatrix}$$

or

$$\mathbf{x}(t)=\mathbf{Q}(t)\mathbf{x}(0)$$

where

$$\mathbf{Q}(t)=\begin{bmatrix}e^{(\sigma+j\omega)t} & 0 & 0\\ 0 & e^{(\sigma-j\omega)t} & 0\\ 0 & 0 & e^{\lambda_3 t}\end{bmatrix}$$

Clearly, $x_1(t)$ and $x_2(t)$ represent an oscillatory motion with its amplitude increasing or decreasing according to $\sigma > 0$ or $\sigma < 0$.

Next we shall consider the case where **A** has multiple eigenvectors. We shall assume that by means of a suitable transformation $\mathbf{x}=\mathbf{P}\mathbf{y}$ the following state space equation:

$$\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$$

is already transformed into

$$\dot{\mathbf{y}}=\mathbf{P}^{-1}\mathbf{A}\mathbf{P}\mathbf{y}=\mathbf{J}\mathbf{y}$$

where **J** is in the Jordan canonical form. Consider the following system:

$$\begin{bmatrix}\dot{y}_1\\ \dot{y}_2\\ \dot{y}_3\end{bmatrix}=\begin{bmatrix}\lambda_1 & 1 & 0\\ 0 & \lambda_1 & 1\\ 0 & 0 & \lambda_1\end{bmatrix}\begin{bmatrix}y_1\\ y_2\\ y_3\end{bmatrix}$$

This is equivalent to the following three equations:

$$\begin{aligned}\dot{y}_1&=\lambda_1 y_1+y_2\\ \dot{y}_2&=\lambda_1 y_2+y_3\\ \dot{y}_3&=\lambda_1 y_3\end{aligned}\tag{3-40}$$

Solving the last of Eq. (3-40) we obtain

$$y_3(t)=e^{\lambda_1 t}y_3(0)\tag{3-41}$$

Substituting Eq. (3-41) into the second part of Eq. (3-40) we obtain

$$\dot{y}_2=\lambda_1 y_2+e^{\lambda_1 t}y_3(0)$$

which yields the solution

$$y_2(t)=e^{\lambda_1 t}y_2(0)+te^{\lambda_1 t}y_3(0)\tag{3-42}$$

Substituting Eq. (3-42) into the first part of Eq. (3-40), we obtain

$$\dot{y}_1=\lambda_1 y_1+e^{\lambda_1 t}y_2(0)+te^{\lambda_1 t}y_3(0)$$

Solving for y_1, we obtain

$$y_1(t)=e^{\lambda_1 t}y_1(0)+te^{\lambda_1 t}y_2(0)+\frac{1}{2}t^2e^{\lambda_1 t}y_3(0)$$

Hence the solution $\mathbf{y}(t)$ can be written as follows:

$$\begin{bmatrix}y_1(t)\\ y_2(t)\\ y_3(t)\end{bmatrix}=\begin{bmatrix}e^{\lambda_1 t} & te^{\lambda_1 t} & \frac{1}{2}t^2e^{\lambda_1 t}\\ 0 & e^{\lambda_1 t} & te^{\lambda_1 t}\\ 0 & 0 & e^{\lambda_1 t}\end{bmatrix}\begin{bmatrix}y_1(0)\\ y_2(0)\\ y_3(0)\end{bmatrix}$$

or

$$\mathbf{y}(t)=\mathbf{S}(t)\mathbf{Q}(t)\mathbf{y}(0)$$

where

$$\mathbf{S}(t)=\begin{bmatrix}1 & t & \frac{1}{2}t^2\\ 0 & 1 & t\\ 0 & 0 & 1\end{bmatrix}$$

and

$$\mathbf{Q}(t)=\begin{bmatrix}e^{\lambda_1 t} & 0 & 0\\ 0 & e^{\lambda_1 t} & 0\\ 0 & 0 & e^{\lambda_1 t}\end{bmatrix}$$

Similarly, we can extend the same analysis to more general cases. As an example consider the following system:

$$\begin{bmatrix}\dot{y}_1\\ \dot{y}_2\\ \dot{y}_3\\ \dot{y}_4\\ \dot{y}_5\\ \dot{y}_6\\ \dot{y}_7\end{bmatrix} = \left[\begin{array}{ccc:cc:cc} \lambda_1 & 1 & 0 & & & & 0 \\ 0 & \lambda_1 & 1 & & & & \\ 0 & 0 & \lambda_1 & & & & \\ \hdashline & & & \lambda_4 & 1 & & \\ & & & 0 & \lambda_4 & & \\ \hdashline & & & & & \lambda_6 & 0 \\ 0 & & & & & 0 & \lambda_7 \end{array}\right] \begin{bmatrix}y_1\\ y_2\\ y_3\\ y_4\\ y_5\\ y_6\\ y_7\end{bmatrix}$$

The solution $\mathbf{y}(t)$ can be written as follows:

$$\begin{bmatrix}y_1(t)\\ y_2(t)\\ y_3(t)\\ y_4(t)\\ y_5(t)\\ y_6(t)\\ y_7(t)\end{bmatrix} = \left[\begin{array}{ccc:cc:cc} e^{\lambda_1 t} & te^{\lambda_1 t} & \frac{1}{2}t^2e^{\lambda_1 t} & & & & 0 \\ 0 & e^{\lambda_1 t} & te^{\lambda_1 t} & & & & \\ 0 & 0 & e^{\lambda_1 t} & & & & \\ \hdashline & & & e^{\lambda_4 t} & te^{\lambda_4 t} & & \\ & & & 0 & e^{\lambda_4 t} & & \\ \hdashline & & & & & e^{\lambda_6 t} & 0 \\ 0 & & & & & 0 & e^{\lambda_7 t} \end{array}\right] \begin{bmatrix}y_1(0)\\ y_2(0)\\ y_3(0)\\ y_4(0)\\ y_5(0)\\ y_6(0)\\ y_7(0)\end{bmatrix}$$

or

$$\mathbf{y}(t) = \mathbf{S}(t)\mathbf{Q}(t)\mathbf{y}(0)$$

where

$$\mathbf{S}(t) = \left[\begin{array}{ccc:cc:cc} 1 & t & \frac{1}{2}t^2 & & & & 0 \\ 0 & 1 & t & & & & \\ 0 & 0 & 1 & & & & \\ \hdashline & & & 1 & t & & \\ & & & 0 & 1 & & \\ \hdashline & & & & & 1 & 0 \\ 0 & & & & & 0 & 1 \end{array}\right]$$

and

$$\mathbf{Q}(t) = \begin{bmatrix} e^{\lambda_1 t} & & & & & & 0 \\ & e^{\lambda_1 t} & & & & & \\ & & e^{\lambda_1 t} & & & & \\ & & & e^{\lambda_4 t} & & & \\ & & & & e^{\lambda_4 t} & & \\ & & & & & e^{\lambda_6 t} & \\ 0 & & & & & & e^{\lambda_7 t} \end{bmatrix}$$

In terms of the original state vector $\mathbf{x}(t)$, the solution is given by

$$\mathbf{x}(t) = \mathbf{P}\mathbf{y}(t) = \mathbf{P}\mathbf{S}(t)\mathbf{Q}(t)\mathbf{P}^{-1}\mathbf{x}(0) \tag{3–43}$$

Notice that Eq. (3–43) is the general solution of Eq. (3–38) if we interpret $\mathbf{S}(t) = \mathbf{I}$ for the case where $\mathbf{A}$ has distinct eigenvectors.

A few comments on motions of free dynamic systems *(3–4)* Consider a system described by $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ where $\mathbf{A}$ is an $n \times n$ constant matrix. Assume that the eigenvalues of $\mathbf{A}$ are $\lambda_1, \lambda_1, \ldots, \lambda_1, \lambda_{m+1}, \ldots, \lambda_n$. In such a system any free motion can be represented uniquely as a linear combination of motions $e^{\lambda_i t}$ $(i = m+1, m+2, \ldots, n)$ and $t^h e^{\lambda_1 t}$ $(h = 0, 1, 2, \ldots, m-1)$. If λ_i is real, the motion $e^{\lambda_i t}$ is along the eigenvector $\mathbf{x}_i$, the eigenvector corresponding to the eigenvalue λ_i. If $\lambda_i > 0$, then the motion $e^{\lambda_i t}$ along the corresponding eigenvector approaches infinity, and if $\lambda_i < 0$, then the motion $e^{\lambda_i t}$ approaches the origin of the state space. If, for example, λ_{m+1} and λ_{m+2} are complex conjugates, however; then the motion $e^{\lambda_{m+1} t}$ and motion $e^{\lambda_{m+2} t}$ cannot be separated from each other. The combined motion takes place in the plane determined by eigenvectors $\mathbf{x}_{m+1}$ and $\mathbf{x}_{m+2}$. The motion in this plane is oscillatory. (Whether the oscillation is damped or growing depends on the sign of the real part of λ_{m+1}.)

The amount of each motion $e^{\lambda_j t}$ $(j = 1, m+1, m+2, \ldots, n)$ depends only on the initial state. (Each motion is excited by the initial condition. For a system with only real λ's, if the initial condition is on the eigenvector $\mathbf{x}_i$ $(i = m+1, m+2, \ldots, n)$, then only the motion $e^{\lambda_i t}$ occurs. If the initial condition is on the multiple eigenvector $\mathbf{x}_1$, only motions $e^{\lambda_1 t}$, $te^{\lambda_1 t}$, $\ldots$, and $t^{m-1}e^{\lambda_1 t}$ will occur. Notice that any initial condition that will excite the motion $e^{\lambda_1 t}$ will also excite the motions $te^{\lambda_1 t}$, $t^2e^{\lambda_1 t}$, $\ldots$, and $t^{m-1}e^{\lambda_1 t}$.

EXERCISE PROBLEMS AND SOLUTIONS

Problem A-3-1 Obtain a transformation matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ is diagonal. $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 3 & 0 & 2 \\ -12 & -7 & -6 \end{bmatrix}$$

Solution The roots of the characteristic equation $|\mathbf{A} - \lambda\mathbf{I}| = 0$ are found from

$$\begin{vmatrix} -\lambda & 1 & 0 \\ 3 & -\lambda & 2 \\ -12 & -7 & -6-\lambda \end{vmatrix} = -(\lambda+1)(\lambda+2)(\lambda+3) = 0$$

as $\lambda_1 = -1$, $\lambda_2 = -2$, and $\lambda_3 = -3$. A diagonalizing transformation matrix $\mathbf{P}$ is given by

$$\mathbf{P} = \begin{bmatrix} \begin{vmatrix} -\lambda_1 & 2 \\ -7 & -6-\lambda_1 \end{vmatrix} & \begin{vmatrix} -\lambda_2 & 2 \\ -7 & -6-\lambda_2 \end{vmatrix} & \begin{vmatrix} -\lambda_3 & 2 \\ -7 & -6-\lambda_3 \end{vmatrix} \\ -\begin{vmatrix} 3 & 2 \\ -12 & -6-\lambda_1 \end{vmatrix} & -\begin{vmatrix} 3 & 2 \\ -12 & -6-\lambda_2 \end{vmatrix} & -\begin{vmatrix} 3 & 2 \\ -12 & -6-\lambda_3 \end{vmatrix} \\ \begin{vmatrix} 3 & -\lambda_1 \\ -12 & -7 \end{vmatrix} & \begin{vmatrix} 3 & -\lambda_2 \\ -12 & -7 \end{vmatrix} & \begin{vmatrix} 3 & -\lambda_3 \\ -12 & -7 \end{vmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} \begin{vmatrix} 1 & 2 \\ -7 & -5 \end{vmatrix} & \begin{vmatrix} 2 & 2 \\ -7 & -4 \end{vmatrix} & \begin{vmatrix} 3 & 2 \\ -7 & -3 \end{vmatrix} \\ -\begin{vmatrix} 3 & 2 \\ -12 & -5 \end{vmatrix} & -\begin{vmatrix} 3 & 2 \\ -12 & -4 \end{vmatrix} & -\begin{vmatrix} 3 & 2 \\ -12 & -3 \end{vmatrix} \\ \begin{vmatrix} 3 & 1 \\ -12 & -7 \end{vmatrix} & \begin{vmatrix} 3 & 2 \\ -12 & -7 \end{vmatrix} & \begin{vmatrix} 3 & 3 \\ -12 & -7 \end{vmatrix} \end{bmatrix} = \begin{bmatrix} 9 & 6 & 5 \\ -9 & -12 & -15 \\ -9 & 3 & 15 \end{bmatrix}$$

We may divide each column of $\mathbf{P}$ by a constant. If we divide the first column by 9, the second column by 3, and the third column by 5, a diagonalizing transformation matrix occurs as

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 2 & 1 \\ -1 & -4 & -3 \\ -1 & 1 & 3 \end{bmatrix}$$

The inverse of $\mathbf{P}_1$ is obtained as

$$\mathbf{P}_1^{-1} = \begin{bmatrix} \frac{9}{2} & \frac{5}{2} & 1 \\ -3 & -2 & -1 \\ \frac{5}{2} & \frac{3}{2} & 1 \end{bmatrix}$$

As a check let us compute $\mathbf{P}_1^{-1}\mathbf{A}\mathbf{P}_1$.

$$\mathbf{P}_1^{-1}\mathbf{A}\mathbf{P}_1 = \begin{bmatrix} \frac{9}{2} & \frac{5}{2} & 1 \\ -3 & -2 & -1 \\ \frac{5}{2} & \frac{3}{2} & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 3 & 0 & 2 \\ -12 & -7 & -6 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ -1 & -4 & -3 \\ -1 & 1 & 3 \end{bmatrix}$$

$$= \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -3 \end{bmatrix}$$

Problem A-3-2 Obtain a transformation matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ is diagonal. $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

Assume that $\mathbf{A}$ has distinct eigenvalues.

Solution Let us denote four distinct eigenvalues as λ_1, λ_2, λ_3, and λ_4, or

$$|\mathbf{A} - \lambda\mathbf{I}| = (\lambda - \lambda_1)(\lambda - \lambda_2)(\lambda - \lambda_3)(\lambda - \lambda_4)$$

The elements in the ith column of $\mathbf{P}$ are given by

$$\mathbf{P}_i = \begin{bmatrix} \begin{vmatrix} a_{22}-\lambda_i & a_{23} & a_{24} \\ a_{32} & a_{33}-\lambda_i & a_{34} \\ a_{42} & a_{43} & a_{44}-\lambda_i \end{vmatrix} \\ -\begin{vmatrix} a_{21} & a_{23} & a_{24} \\ a_{31} & a_{33}-\lambda_i & a_{34} \\ a_{41} & a_{43} & a_{44}-\lambda_i \end{vmatrix} \\ \begin{vmatrix} a_{21} & a_{22}-\lambda_i & a_{24} \\ a_{31} & a_{32} & a_{34} \\ a_{41} & a_{42} & a_{44}-\lambda_i \end{vmatrix} \\ -\begin{vmatrix} a_{21} & a_{22}-\lambda_i & a_{23} \\ a_{31} & a_{32} & a_{33}-\lambda_i \\ a_{41} & a_{42} & a_{43} \end{vmatrix} \end{bmatrix}$$

$(i = 1, 2, 3, 4)$

Problem A-3-3 Prove that any $n \times n$ real symmetric matrix $\mathbf{A}$ is orthogonally similar to a diagonal matrix.

Solution Let us denote the eigenvalues of $\mathbf{A}$ as $\lambda_1, \lambda_2, \ldots, \lambda_n$. The λ_i are real. There exists a normalized real eigenvector $\mathbf{x}_1$ satisfying

$$\mathbf{A}\mathbf{x}_1 = \lambda_1\mathbf{x}_1$$

Let $\mathbf{T}_1$ be an orthogonal matrix with $\mathbf{x}_1$ as its first column. Then the first column of $\mathbf{T}_1^{-1}\mathbf{A}\mathbf{T}_1$ is obtained as follows (refer to Theorem 3-4):

$$\begin{bmatrix} \lambda_1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

Hence noting that $\mathbf{T}_1$ is orthogonal and $\mathbf{A}$ is real symmetric, we may write $\mathbf{T}_1^{-1}\mathbf{A}\mathbf{T}_1$ as follows:

$$\mathbf{T}_1^{-1}\mathbf{A}\mathbf{T}_1 = \left[\begin{array}{c|c} \lambda_1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{A}_1 \end{array}\right]$$

where $\mathbf{A}_1 = (n-1) \times (n-1)$ real symmetric matrix whose eigenvlaues are $\lambda_2, \lambda_3, \ldots, \lambda_n$

$\mathbf{0}$ = matrix consisting of zeros

Similarly, we can find an $(n-1) \times (n-1)$ orthogonal matrix $\mathbf{S}$ such that

$$\mathbf{S}^{-1}\mathbf{A}_1\mathbf{S} = \left[\begin{array}{c|c} \lambda_2 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{A}_2 \end{array}\right]$$

where $\mathbf{A}_2 = (n-2) \times (n-2)$ real symmetric matrix whose eigenvalues are $\lambda_3, \lambda_4, \ldots, \lambda_n$

By letting

$$\mathbf{T}_2 = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & \mathbf{S} \end{array}\right]$$

we obtain

$$\mathbf{T}_2^{-1}(\mathbf{T}_1^{-1}\mathbf{A}\mathbf{T}_1)\mathbf{T}_2 = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & \mathbf{S}^{-1} \end{array}\right]\left[\begin{array}{c|c} \lambda_1 & 0 \\ \hline 0 & \mathbf{A}_1 \end{array}\right]\left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & \mathbf{S} \end{array}\right]$$

$$= \left[\begin{array}{c|c} \lambda_1 & 0 \\ \hline 0 & \mathbf{S}^{-1}\mathbf{A}_1\mathbf{S} \end{array}\right] = \left[\begin{array}{cc|c} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & \\ \hline & 0 & \mathbf{A}_2 \end{array}\right]$$

Repeating the same process we can finally obtain

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \mathbf{T}'\mathbf{A}\mathbf{T} = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix}$$

where $\mathbf{T} = \mathbf{T}_1\mathbf{T}_2 \cdots \mathbf{T}_n$

Problem A-3-4 Suppose that an $n \times n$ normal matrix $\mathbf{A}$ possesses real eigenvalues. Prove that $\mathbf{A}$ is Hermitian.

Solution Since $\mathbf{A}$ is normal, there exists a unitary matrix $\mathbf{U}$ such that

$$\mathbf{U}^{-1}\mathbf{A}\mathbf{U} = \mathbf{D} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are real. Then

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^{-1} = \mathbf{U}\mathbf{D}\mathbf{U}^* = \mathbf{U}\mathbf{D}^*\mathbf{U}^* = \mathbf{A}^*$$

Hence $\mathbf{A}$ is Hermitian.

Problem A-3-5 Find a transformation matrix $\mathbf{S}$ such that

$$\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{J}$$

where

$$\mathbf{A} = \begin{bmatrix} 4 & 1 & -2 \\ 1 & 0 & 2 \\ 1 & -1 & 3 \end{bmatrix}, \qquad \mathbf{J} = \begin{bmatrix} 3 & 1 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Solution The characteristic equation becomes

$$|\mathbf{A} - \lambda\mathbf{I}| = -\lambda^3 + 7\lambda^2 - 15\lambda + 9 = -(\lambda - 3)(\lambda - 3)(\lambda - 1)$$

Hence the eigenvalues of $\mathbf{A}$ are 3, 3, 1. Simple computation reveals that there exist two linearly independent eigenvectors. Therefore, it is possible to transform $\mathbf{A}$ to $\mathbf{J}$ by a suitable matrix $\mathbf{S}$. Such a matrix can be obtained from Eq. (3–26). We shall denote this as $\mathbf{S}_1$.

$$\mathbf{S}_1 = \begin{bmatrix} \begin{vmatrix} -\lambda_1 & 2 \\ -1 & 3-\lambda_1 \end{vmatrix} & 2\lambda_1 - 3 & \begin{vmatrix} -\lambda_3 & 2 \\ -1 & 3-\lambda_3 \end{vmatrix} \\ -\begin{vmatrix} 1 & 2 \\ 1 & 3-\lambda_1 \end{vmatrix} & 1 & -\begin{vmatrix} 1 & 2 \\ 1 & 3-\lambda_3 \end{vmatrix} \\ \begin{vmatrix} 1 & -\lambda_1 \\ 1 & -1 \end{vmatrix} & 1 & \begin{vmatrix} 1 & -\lambda_3 \\ 1 & -1 \end{vmatrix} \end{bmatrix}$$

By substituting $\lambda_1 = 3$, $\lambda_3 = 1$ into $\mathbf{S}_1$, we obtain

$$\mathbf{S}_1 = \begin{bmatrix} 2 & 3 & 0 \\ 2 & 1 & 0 \\ 2 & 1 & 0 \end{bmatrix}$$

Notice that the elements of the third column are zeros. Therefore, we must use other than $A_{311}, A_{312}, A_{313}$ for the third column. Let us take $A_{321}, A_{322}, A_{323}$ where the A_{32j} are the cofactors of the second row of the determinant

$$\begin{vmatrix} a-\lambda_3 & b & c \\ d & e-\lambda_3 & f \\ g & h & i-\lambda_3 \end{vmatrix}$$

Namely,

$$A_{321} = -\begin{vmatrix} b & c \\ h & i-\lambda_3 \end{vmatrix}, \qquad A_{322} = \begin{vmatrix} a-\lambda_3 & c \\ g & i-\lambda_3 \end{vmatrix}, \qquad A_{323} = -\begin{vmatrix} a-\lambda_3 & b \\ g & h \end{vmatrix}$$

Then $\mathbf{S}$ becomes as follows:

$$\mathbf{S} = \begin{bmatrix} \begin{vmatrix} -\lambda_1 & 2 \\ -1 & 3-\lambda_1 \end{vmatrix} & 2\lambda_1 - 3 & -\begin{vmatrix} 1 & -2 \\ -1 & 3-\lambda_3 \end{vmatrix} \\ -\begin{vmatrix} 1 & 2 \\ 1 & 3-\lambda_1 \end{vmatrix} & 1 & \begin{vmatrix} 4-\lambda_3 & -2 \\ 1 & 3-\lambda_3 \end{vmatrix} \\ \begin{vmatrix} 1 & -\lambda_1 \\ 1 & -1 \end{vmatrix} & 1 & -\begin{vmatrix} 4-\lambda_3 & 1 \\ 1 & -1 \end{vmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 3 & 0 \\ 2 & 1 & 8 \\ 2 & 1 & 4 \end{bmatrix}$$

By substitution, we can easily verify that $\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{J}$.

Notice that $\mathbf{S}$ is not unique. If we postmultiply $\mathbf{S}$ by $\mathbf{V}$, where, for example,

$$\mathbf{V} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix}$$

then

$$\mathbf{S}\mathbf{V} = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 2 \\ 1 & 1 & 1 \end{bmatrix}$$

will satisfy $(\mathbf{SV})^{-1}\mathbf{A}(\mathbf{SV}) = \mathbf{J}$. Hence examples of transformation matrices that transform $\mathbf{A}$ to $\mathbf{J}$ are

$$\begin{bmatrix} 2 & 3 & 0 \\ 2 & 1 & 8 \\ 2 & 1 & 4 \end{bmatrix}, \quad \begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 2 \\ 1 & 1 & 1 \end{bmatrix}, \quad \text{etc.}$$

Problem A-3-6 Given the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_2 & -a_1 \end{bmatrix}$$

show that the cofactors A_{i1j} $(i, j = 1, 2, \ldots, n)$ of the determinant formed as

$$|\mathbf{A} - \lambda_i \mathbf{I}|$$

are

$$\begin{aligned} A_{i11} &= (-)^n \frac{a_n}{\lambda_i} \\ A_{i12} &= (-)^n a_n \\ A_{i13} &= (-)^n a_n \lambda_i \\ &\cdots \\ A_{i1n} &= (-)^n a_n \lambda_i^{n-2} \\ (i &= 1, 2, \ldots, n) \end{aligned} \tag{3-44}$$

Assuming the eigenvalues of $\mathbf{A}$ are distinct, show that the transformation matrix defined by (3-21) that transforms $\mathbf{A}$ into a diagonal matrix can be given by

$$\mathbf{P} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_n \\ \lambda_1^2 & \lambda_2^2 & \cdots & \lambda_n^2 \\ \vdots & \vdots & & \vdots \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \cdots & \lambda_n^{n-1} \end{bmatrix}$$

Solution

$$|\mathbf{A} - \lambda_i \mathbf{I}| = \begin{vmatrix} -\lambda_i & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\lambda_i & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda_i & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_2 & -a_1 - \lambda_i \end{vmatrix}$$

The sum of the products of the elements of any row and the cofactors of corresponding elements of a different row is zero. Thus, for the elements of the second row and cofactors of the first row:

$$0A_{i11} - \lambda_i A_{i12} + A_{i13} + 0A_{i14} + \cdots + 0A_{i1n} = 0$$

Thus,

$$A_{i13} = \lambda_i A_{i12}$$

For the elements of the jth row $(j = 2, 3, \ldots, n-1)$ and cofactors of the first row:

$$0A_{i11} + \cdots + 0A_{i1(j-1)} - \lambda_i A_{i1j} + A_{i1(j+1)} + 0A_{i1(j+2)} + \cdots + 0\,A_{i1n} = 0$$

Thus,

$$A_{i1(j+1)} = \lambda_i A_{i1j} \tag{3-45}$$

For $j = n - 1$ we obtain

$$A_{i1n} = \lambda_i A_{i1(n-1)}$$

A_{i1n} can be computed as follows:

$$A_{i1n} = (-)^{n+1} \begin{vmatrix} 0 & -\lambda_i & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -\lambda_i & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & . & . & . & \cdots & 0 & -\lambda_i \\ -a_n & -a_{n-1} & . & . & \cdots & -a_3 & -a_2 \end{vmatrix}$$

Expanding by minors about the first column

$$A_{i1n} = (-)^{n+1}(-)^{1+(n-1)}(-a_n) \begin{vmatrix} -\lambda_i & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\lambda_i & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & -\lambda_i \end{vmatrix}$$

Again expanding by minors about the first column

$$A_{i1n} = (-)^{2n+2} a_n (-\lambda_i) \begin{vmatrix} -\lambda_i & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\lambda_i & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & -\lambda_i \end{vmatrix}$$

It can be easily seen that by continuing the expansion about the first column the result is

$$A_{i1n} = (-)^{2n+2} a_n (-\lambda_i)(-\lambda_i)^{n-3} = (-)^n a_n \lambda_i^{n-2}$$

This proves the last of Eq. (3-44). Then using Eq. (3-45), rearranged

$$A_{i1j} = \frac{1}{\lambda_i} A_{i1(j+1)}$$

we obtain

$$\begin{aligned} A_{i11} &= \frac{1}{\lambda_i} A_{i12} = \frac{1}{\lambda_i^2} A_{i13} = \frac{1}{\lambda_i^{n-1}} A_{i1n} = (-)^n a_n \frac{1}{\lambda_i} \\ A_{i12} &= (-)^n a_n \\ A_{i13} &= (-)^n a_n \lambda_i \\ &\cdots \end{aligned}$$

The transformation matrix $\mathbf{P}$ that transforms $\mathbf{A}$ into the diagonal form can be obtained as follows: By dividing each column of the following matrix (A_{i1j}), the matrix defined by Eq. (3-18),

$$\begin{bmatrix} (-)^n a_n \frac{1}{\lambda_1} & (-)^n a_n \frac{1}{\lambda_2} & \cdots & (-)^n a_n \frac{1}{\lambda_n} \\ (-)^n a_n & (-)^n a_n & \cdots & (-)^n a_n \\ \vdots & \vdots & & \vdots \\ (-)^n a_n \lambda_1^{n-2} & (-)^n a_n \lambda_2^{n-2} & \cdots & (-)^n a_n \lambda_n^{n-2} \end{bmatrix}$$

by $(-)^n a_n (1/\lambda_i)$ we obtain $\mathbf{P}$ in the following form:

$$\mathbf{P} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_n \\ \lambda_1^2 & \lambda_2^2 & \cdots & \lambda_n^2 \\ \vdots & \vdots & & \vdots \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \cdots & \lambda_n^{n-1} \end{bmatrix}$$

Problem A-3-7 Consider the system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_{n-1} \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_2 & -a_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} \tag{3-46}$$

Suppose that the characteristic equation involves a pair of double roots. Derive a transformation matrix $\mathbf{S}$ such that after the coordinates are transformed the equation is in the Jordan canonical form.

Next obtain a transformation matrix that will transform the system equation into the Jordan canonical form when the characteristic equation involves multiple roots of multiplicity four.

Solution We may first assume that the double roots are slightly different from each other and write the roots of the characteristic equation as λ_1, $\lambda_1 + \Delta\lambda_1, \lambda_3, \ldots, \lambda_n$ and obtain a transformation matrix $\mathbf{S}$ and then let $\Delta\lambda_1 \longrightarrow 0$. If we denote the transformation matrix as obtained in Prob. A-3-6 as $\mathbf{P}$, then

$$\mathbf{P} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_1 + \Delta\lambda_1 & \lambda_3 & \cdots & \lambda_n \\ \vdots & \vdots & \vdots & & \vdots \\ \lambda_1^{n-1} & (\lambda_1 + \Delta\lambda_1)^{n-1} & \lambda_3^{n-1} & \cdots & \lambda_n^{n-1} \end{bmatrix} \tag{3-47}$$

After the transformation $\mathbf{x} = \mathbf{P}\mathbf{y}$ is made, the system equations become

$$\begin{aligned} \dot{y}_1 &= \lambda_1 y_1 \\ \dot{y}_2 &= (\lambda_1 + \Delta\lambda_1) y_2 \\ \dot{y}_3 &= \lambda_3 y_3 \\ &\cdots \\ \dot{y}_n &= \lambda_n y_n \end{aligned}$$

If we perform the second transformation given by $\mathbf{y} = \mathbf{Q}\mathbf{z}$, where

$$\mathbf{Q} = \begin{bmatrix} 0 & -\frac{1}{\Delta\lambda_1} & 0 & \cdots & 0 \\ 1 & \frac{1}{\Delta\lambda_1} & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

then we obtain the following equations:

$$\begin{aligned} \dot{z}_1 &= (\lambda_1 + \Delta\lambda_1) z_1 + z_2 \\ \dot{z}_2 &= \lambda_1 z_2 \\ \dot{z}_3 &= \lambda_3 z_3 \\ &\cdots \\ \dot{z}_n &= \lambda_n z_n \end{aligned} \tag{3-48}$$

Notice that although $\mathbf{P}$ becomes singular as $\Delta\lambda_1 \longrightarrow 0$, $\mathbf{PQ}$ is nonsingular as $\Delta\lambda_1 \longrightarrow 0$. Letting $\Delta\lambda_1$ approach zero in Eq. (3-48) we obtain

$$\begin{aligned} \dot{z}_1 &= \lambda_1 z_1 + z_2 \\ \dot{z}_2 &= \lambda_1 z_2 \\ \dot{z}_3 &= \lambda_3 z_3 \\ &\cdots \\ \dot{z}_n &= \lambda_n z_n \end{aligned} \tag{3-49}$$

Equations (3-49) yield the Jordan canonical form when the characteristic equation of Eq. (3-46) involves double eigenvalues $\lambda_1 = \lambda_2$. The transformation matrix $\mathbf{S}$ that transforms Eq. (3-46) into Eq. (3-49) can be obtained as follows:

$$\begin{aligned} \mathbf{S} &= \lim_{\Delta\lambda_1 \to 0} \mathbf{PQ} \\ &= \lim_{\Delta\lambda_1 \to 0} \begin{bmatrix} 1 & 0 & 1 & \cdots & 1 \\ \lambda_1 + \Delta\lambda_1 & 1 & \lambda_3 & \cdots & \lambda_n \\ (\lambda_1 + \Delta\lambda_1)^2 & 2\lambda_1 + \Delta\lambda_1 & \lambda_3^2 & \cdots & \lambda_n^2 \\ (\lambda_1 + \Delta\lambda_1)^3 & 3\lambda_1^2 + 3\lambda_1\Delta\lambda_1 + \Delta\lambda_1^2 & \lambda_3^3 & \cdots & \lambda_n^3 \\ (\lambda_1 + \Delta\lambda_1)^4 & 4\lambda_1^3 + 6\lambda_1^2\Delta\lambda_1 + 4\lambda_1\Delta\lambda_1^2 + \Delta\lambda_1^3 & \lambda_3^4 & \cdots & \lambda_n^4 \\ \vdots & \vdots & \vdots & & \vdots \\ (\lambda_1 + \Delta\lambda_1)^{n-1} & \cdots & \lambda_3^{n-1} & \cdots & \lambda_n^{n-1} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 1 & \cdots & 1 \\ \lambda_1 & 1 & \lambda_3 & \cdots & \lambda_n \\ \lambda_1^2 & 2\lambda_1 & \lambda_3^2 & \cdots & \lambda_n^2 \\ \lambda_1^3 & 3\lambda_1^2 & \lambda_3^3 & \cdots & \lambda_n^3 \\ \vdots & \vdots & \vdots & & \vdots \\ \lambda_1^{n-1} & (n-1)\lambda_1^{n-2} & \lambda_3^{n-1} & \cdots & \lambda_n^{n-1} \end{bmatrix} \end{aligned} \tag{3-50}$$

Notice that the elements in the second column are derivatives (with respect to λ_1)

of the corresponding elements in the first column. After the coordinates are transformed, the equation in z coordinates becomes as follows:

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \\ \vdots \\ \dot{z}_n \end{bmatrix} = \begin{bmatrix} \lambda_1 & 1 & 0 & \cdots & 0 \\ 0 & \lambda_1 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_n \end{bmatrix} \tag{3-51}$$

If the characteristic equation involves multiple roots of multiplicity four, then the following transformation matrix:

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & \cdots & 1 \\ \lambda_1 & 1 & 0 & 0 & \lambda_5 & \cdots & \lambda_n \\ \lambda_1^2 & 2\lambda_1 & 1 & 0 & \lambda_5^2 & \cdots & \lambda_n^2 \\ \lambda_1^3 & 3\lambda_1^2 & 3\lambda_1 & 1 & \lambda_5^3 & \cdots & \lambda_n^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ \lambda_1^{n-1} & \dfrac{d}{d\lambda_1}(\lambda_1^{n-1}) & \dfrac{1}{2!}\dfrac{d^2}{d\lambda_1^2}(\lambda_1^{n-1}) & \dfrac{1}{3!}\dfrac{d^3}{d\lambda_1^3}(\lambda_1^{n-1}) & \lambda_5^{n-1} & \cdots & \lambda_n^{n-1} \end{bmatrix} \tag{3-52}$$

will transform Eq. (3–46) into the following form:

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \\ \dot{z}_4 \\ \dot{z}_5 \\ \vdots \\ \dot{z}_n \end{bmatrix} = \left[\begin{array}{cccc:ccc} \lambda_1 & 1 & 0 & 0 & & & 0 \\ 0 & \lambda_1 & 1 & 0 & & & \\ 0 & 0 & \lambda_1 & 1 & & & \\ 0 & 0 & 0 & \lambda_1 & & & \\ \hdashline & & & & \lambda_5 & & 0 \\ & & & & & \ddots & \\ 0 & & & & 0 & & \lambda_n \end{array}\right] \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ \vdots \\ z_n \end{bmatrix}$$

Notice that in Eq. (3–52) the second column is the first derivative (with respect to λ_1) of the first column, the third column is the second derivative of the first column divided by 2!, and the fourth column is the third derivative of the first column divided by 3!.

Problem A-3-8 Consider the differential equation

$$\dddot{x} + a_1\ddot{x} + a_2\dot{x} + a_3x = 0$$

or its state space representation

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_3 & -a_2 & -a_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Assume that the characteristic equation includes two equal eigenvalues, or the eigenvalues are $\lambda_1, \lambda_1, \lambda_3$. Show directly that the following transformation

$$\mathbf{x} = \mathbf{P}\mathbf{y}$$

where

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & \lambda_1 & \lambda_3 \\ 2\lambda_1 & \lambda_1^2 & \lambda_3^2 \end{bmatrix}$$

will reduce the original equation into the following form:

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 1 & \lambda_1 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

Solution We shall denote the coefficient matrix of the given state space equation as **A**.

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_3 & -a_2 & -a_1 \end{bmatrix}$$

By the transformation $\mathbf{x} = \mathbf{P}\mathbf{y}$, we obtain

$$\mathbf{P}\dot{\mathbf{y}} = \mathbf{A}\mathbf{P}\mathbf{y}$$

Forming the product **AP**,

$$\mathbf{AP} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_3 & -a_2 & -a_1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 \\ 1 & \lambda_1 & \lambda_3 \\ 2\lambda_1 & \lambda_1^2 & \lambda_3^2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \lambda_1 & \lambda_3 \\ 2\lambda_1 & \lambda_1^2 & \lambda_3^2 \\ -a_2 - 2a_1\lambda_1 & -a_3 - a_2\lambda_1 - a_1\lambda_1^2 & -a_3 - a_2\lambda_3 - a_1\lambda_3^2 \end{bmatrix}$$

Since the eigenvalues are $\lambda_1, \lambda_1, \lambda_3$, the characteristic equation becomes

$$(\lambda - \lambda_1)(\lambda - \lambda_1)(\lambda - \lambda_3) = \lambda^3 + (-2\lambda_1 - \lambda_3)\lambda^2 + (\lambda_1^2 + 2\lambda_1\lambda_3)\lambda + (-\lambda_1^2\lambda_3) = 0$$

Hence we obtain

$$a_1 = -2\lambda_1 - \lambda_3$$
$$a_2 = \lambda_1^2 + 2\lambda_1\lambda_3$$
$$a_3 = -\lambda_1^2\lambda_3$$

The matrix **AP** can then be written as

$$\mathbf{AP} = \begin{bmatrix} 1 & \lambda_1 & \lambda_3 \\ 2\lambda_1 & \lambda_1^2 & \lambda_3^2 \\ 3\lambda_1^2 & \lambda_1^3 & \lambda_3^3 \end{bmatrix}$$

Since $\mathbf{P}^{-1}$ is obtained as

$$\mathbf{P}^{-1} = \frac{1}{\lambda_1^2 - 2\lambda_1\lambda_3 + \lambda_3^2} \begin{bmatrix} \lambda_1^2\lambda_3 - \lambda_3^2\lambda_1 & \lambda_3^2 - \lambda_1^2 & \lambda_1 - \lambda_3 \\ \lambda_3^2 - 2\lambda_1\lambda_3 & 2\lambda_1 & -1 \\ \lambda_1^2 & -2\lambda_1 & 1 \end{bmatrix}$$

we obtain

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 1 & \lambda_1 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$$

Notice that

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & 0 \\ 1 & \lambda_1 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 1 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$$

Problem A-3-9 Simplify the following state space equation:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -a_4 & -a_3 & -a_2 & -a_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

where the characteristic equation

$$\lambda^4 + a_1\lambda^3 + a_2\lambda^2 + a_3\lambda + a_4 = 0$$

is assumed to possess two identical roots and one pair of complex conjugate roots, that is, $\lambda_1 = \lambda_2$, $\lambda_3 = \sigma + j\omega$, $\lambda_4 = \sigma - j\omega$.

Solution The following transformation

$$\mathbf{x} = \mathbf{P}\mathbf{y}$$

where

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 1 & 1 \\ \lambda_1 & 1 & \sigma + j\omega & \sigma - j\omega \\ \lambda_1^2 & 2\lambda_1 & (\sigma + j\omega)^2 & (\sigma - j\omega)^2 \\ \lambda_1^3 & 3\lambda_1^2 & (\sigma + j\omega)^3 & (\sigma - j\omega)^3 \end{bmatrix}$$

will transform the given state space equation into the following form:

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \\ \dot{y}_4 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & \sigma + j\omega & 0 \\ 0 & 0 & 0 & \sigma - j\omega \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

Let us carry out the second transformation

$$\mathbf{y} = \mathbf{Q}\mathbf{z}$$

where

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & -\frac{j}{2} \\ 0 & 0 & \frac{1}{2} & \frac{j}{2} \end{bmatrix}$$

then the transformed equation becomes

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \\ \dot{z}_4 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & \sigma & \omega \\ 0 & 0 & -\omega & \sigma \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix}$$

Problem A-3-10 Given the following state space equation:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} -a_1 & 1 & 0 & 0 & 0 \\ -a_2 & 0 & 1 & 0 & 0 \\ -a_3 & 0 & 0 & 1 & 0 \\ -a_4 & 0 & 0 & 0 & 1 \\ -a_5 & 0 & 0 & 0 & 0 \end{bmatrix}$$

obtain the transformation matrix $\mathbf{P}$ defined by Eq. (3–18) that will transform the coefficient matrix into the diagonal form. Assume the eigenvalues of $\mathbf{A}$ are distinct.

Solution The (i, j)th element of $\mathbf{P}$ is A_{ij}. In order to determine A_{ij}, let us form the characteristic equation and substitute $\lambda = \lambda_i$ in the equation.

$$|\mathbf{A} - \lambda_i\mathbf{I}| = \begin{vmatrix} -a_1 - \lambda_i & 1 & 0 & 0 & 0 \\ -a_2 & -\lambda_i & 1 & 0 & 0 \\ -a_3 & 0 & -\lambda_i & 1 & 0 \\ -a_4 & 0 & 0 & -\lambda_i & 1 \\ -a_5 & 0 & 0 & 0 & -\lambda_i \end{vmatrix} = 0$$

Since the sum of the products of the elements of any row and the cofactors of corresponding elements of a different row is zero, we obtain for the elements of the last row and cofactors of the first row:

$$-a_5A_{i11} - \lambda_iA_{i15} = 0$$

Thus,

$$A_{i15} = -\frac{a_5}{\lambda_i}A_{i11}$$

For the elements of the fourth row and cofactors of the first row:

$$-a_4A_{i11} - \lambda_iA_{i14} + A_{i15} = 0$$

Thus,

$$A_{i14} = \left(-\frac{a_4}{\lambda_i} - \frac{a_5}{\lambda_i^2}\right)A_{i11}$$

Similarly, we obtain

$$A_{i13} = \left(-\frac{a_3}{\lambda_i} - \frac{a_4}{\lambda_i^2} - \frac{a_5}{\lambda_i^3}\right)A_{i11}$$

$$A_{i12} = \left(-\frac{a_2}{\lambda_i} - \frac{a_3}{\lambda_i^2} - \frac{a_4}{\lambda_i^3} - \frac{a_5}{\lambda_i^4}\right)A_{i11}$$

A_{i11} is obtained as

$$A_{i11} = \lambda_i^4$$

Hence we obtain

$$\begin{aligned} A_{i11} &= \lambda_i^4 \\ A_{i12} &= -(a_2\lambda_i^3 + a_3\lambda_i^2 + a_4\lambda_i + a_5) \\ A_{i13} &= -(a_3\lambda_i^3 + a_4\lambda_i^2 + a_5\lambda_i) \\ A_{i14} &= -(a_4\lambda_i^3 + a_5\lambda_i^2) \\ A_{i15} &= -a_5\lambda_i^3 \end{aligned}$$

Then $\mathbf{P}$ can be given by

$$\mathbf{P} = \begin{bmatrix} \lambda_1^4 & \cdots & \lambda_5^4 \\ -(a_2\lambda_1^3 + a_3\lambda_1^2 + a_4\lambda_1 + a_5) & \cdots & -(a_2\lambda_5^3 + a_3\lambda_5^2 + a_4\lambda_5 + a_5) \\ -(a_3\lambda_1^3 + a_4\lambda_1^2 + a_5\lambda_1) & \cdots & -(a_3\lambda_5^3 + a_4\lambda_5^2 + a_5\lambda_5) \\ -(a_4\lambda_1^3 + a_5\lambda_1^2) & \cdots & -(a_4\lambda_5^3 + a_5\lambda_5^2) \\ -a_5\lambda_1^3 & \cdots & -a_5\lambda_5^3 \end{bmatrix}$$

Notice that since

$$\lambda_i^5 + a_1\lambda_i^4 + a_2\lambda_i^3 + a_3\lambda_i^2 + a_4\lambda_i + a_5 = 0$$

we have

$$\begin{aligned} -(a_2\lambda_i^3 + a_3\lambda_i^2 + a_4\lambda_i + a_5) &= \lambda_i^5 + a_1\lambda_i^4 \\ -(a_3\lambda_i^3 + a_4\lambda_i^2 + a_5\lambda_i) &= \lambda_i^6 + a_1\lambda_i^5 + a_2\lambda_i^4 \\ -(a_4\lambda_i^3 + a_5\lambda_i^2) &= \lambda_i^7 + a_1\lambda_i^6 + a_2\lambda_i^5 + a_3\lambda_i^4 \\ -a_5\lambda_i^3 &= \lambda_i^8 + a_1\lambda_i^7 + a_2\lambda_i^6 + a_3\lambda_i^5 + a_4\lambda_i^4 \end{aligned}$$

Then, by dividing each column of $\mathbf{P}$ by λ_i^4 we obtain a transformation matrix $\mathbf{P}_1$ as follows:

$$\mathbf{P}_1 = \begin{bmatrix} 1 & \cdots & 1 \\ \lambda_1 + a_1 & \cdots & \lambda_5 + a_1 \\ \lambda_1^2 + a_1\lambda_1 + a_2 & \cdots & \lambda_5^2 + a_1\lambda_5 + a_2 \\ \lambda_1^3 + a_1\lambda_1^2 + a_2\lambda_1 + a_3 & \cdots & \lambda_5^3 + a_1\lambda_5^2 + a_2\lambda_5 + a_3 \\ \lambda_1^4 + a_1\lambda_1^3 + a_2\lambda_1^2 + a_3\lambda_1 + a_4 & \cdots & \lambda_5^4 + a_1\lambda_5^3 + a_2\lambda_5^2 + a_3\lambda_5 + a_4 \end{bmatrix}$$

Problem A-3-11 The transformation matrix $\mathbf{P}$ defined by

$$\mathbf{P} = \begin{bmatrix} 1 & 1 & 1 \\ \lambda_1 + a_1 & \lambda_2 + a_1 & \lambda_3 + a_1 \\ \lambda_1^2 + a_1\lambda_1 + a_2 & \lambda_2^2 + a_1\lambda_2 + a_2 & \lambda_3^2 + a_1\lambda_3 + a_2 \end{bmatrix}$$

will transform the coefficient matrix of the following state space equation into the diagonal form:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} -a_1 & 1 & 0 \\ -a_2 & 0 & 1 \\ -a_3 & 0 & 0 \end{bmatrix}$$

$$|\mathbf{A} - \lambda\mathbf{I}| = -(\lambda^3 + a_1\lambda^2 + a_2\lambda + a_3) = -(\lambda - \lambda_1)(\lambda - \lambda_2)(\lambda - \lambda_3)$$

provided that λ_1, λ_2, and λ_3 are distinct. Show that if $\lambda_1 = \lambda_2 \neq \lambda_3$, the following

matrix $\mathbf{S}$ will transform the coefficient matrix into the Jordan canonical form:

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 1 \\ \lambda_1 + a_1 & 1 & \lambda_3 + a_1 \\ \lambda_1^2 + a_1\lambda_1 + a_2 & 2\lambda_1 + a_1 & \lambda_3^2 + a_1\lambda_3 + a_2 \end{bmatrix}$$

Show also that if $\lambda_1 = \lambda_2 = \lambda_3$, $\mathbf{S}$ must be modified into $\mathbf{T}$ where

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_1 + a_1 & 1 & 0 \\ \lambda_1^2 + a_1\lambda_1 + a_2 & 2\lambda_1 + a_1 & 1 \end{bmatrix} \tag{3-53}$$

Solution

Case 1: where $\lambda_1 = \lambda_2 \neq \lambda_3$ The transformation

$$\mathbf{x} = \mathbf{P}_1\mathbf{y}$$

where

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 1 & 1 \\ \lambda_1 + a_1 & \lambda_1 + \Delta\lambda_1 + a_1 & \lambda_3 + a_1 \\ \lambda_1^2 + a_1\lambda_1 + a_2 & (\lambda_1 + \Delta\lambda_1)^2 + a_1(\lambda_1 + \Delta\lambda_1) + a_2 & \lambda_3^2 + a_1\lambda_3 + a_2 \end{bmatrix}$$

will transform the original state space equation into the following form:

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_1 + \Delta\lambda_1 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

If we perform the second transformation

$$\mathbf{y} = \mathbf{Q}_1\mathbf{z}$$

where

$$\mathbf{Q}_1 = \begin{bmatrix} 0 & -\dfrac{1}{\Delta\lambda_1} & 0 \\ 1 & \dfrac{1}{\Delta\lambda_1} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

then we obtain

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 + \Delta\lambda_1 & 1 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

Noting that $\mathbf{P}_1\,\mathbf{Q}_1$ is nonsingular as $\Delta\lambda_1 \longrightarrow 0$ we see that by the transformation

$$\mathbf{x} = \mathbf{S}\mathbf{z} = \lim_{\Delta\lambda_1 \to 0} \mathbf{P}_1\mathbf{Q}_1\mathbf{z}$$

we can transform the coefficient matrix of the original state space equation into the Jordan canonical form. The transformation matrix $\mathbf{S}$ is given by

$$\mathbf{S} = \lim_{\Delta\lambda_1 \to 0} \mathbf{P}_1\mathbf{Q}_1 = \begin{bmatrix} 1 & 0 & 1 \\ \lambda_1 + a_1 & 1 & \lambda_3 + a_1 \\ \lambda_1^2 + a_1\lambda_1 + a_2 & 2\lambda_1 + a_1 & \lambda_3^2 + a_1\lambda_3 + a_2 \end{bmatrix}$$

Case 2: where $\lambda_1 = \lambda_2 = \lambda_3$ The transformation

$$\mathbf{x} = \mathbf{P}_2\mathbf{y}$$

where

$$\mathbf{P}_2 = \begin{bmatrix} 1 & 1 & 1 \\ \lambda_1 + a_1 & \lambda_1 + \Delta\lambda_1 + a_1 & \lambda_1 + 2\Delta\lambda_1 + a_1 \\ \lambda_1^2 + a_1\lambda_1 + a_2 & (\lambda_1 + \Delta\lambda_1)^2 + a_1(\lambda_1 + \Delta\lambda_1) + a_2 & (\lambda_1 + 2\Delta\lambda_1)^2 + a_1(\lambda_1 + 2\Delta\lambda_1) + a_2 \end{bmatrix}$$

will transform the original state space equation into the following form:

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_1 + \Delta\lambda_1 & 0 \\ 0 & 0 & \lambda_1 + 2\Delta\lambda_1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

If we perform the second transformation

$$\mathbf{y} = \mathbf{Q}_2\mathbf{z}$$

where

$$\mathbf{Q}_2 = \begin{bmatrix} 0 & 0 & \dfrac{1}{2\Delta\lambda_1^2} \\ 0 & -\dfrac{1}{\Delta\lambda_1} & -\dfrac{1}{\Delta\lambda_1^2} \\ 1 & \dfrac{1}{\Delta\lambda_1} & \dfrac{1}{2\Delta\lambda_1^2} \end{bmatrix}$$

then we obtain

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 + 2\Delta\lambda_1 & 1 & 0 \\ 0 & \lambda_1 + \Delta\lambda_1 & 1 \\ 0 & 0 & \lambda_1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

Noting that $\mathbf{P}_2\,\mathbf{Q}_2$ is nonsingular as $\Delta\lambda_1 \longrightarrow 0$ we see that by the transformation

$$\mathbf{x} = \mathbf{T}\mathbf{z} = \lim_{\Delta\lambda_1 \to 0} \mathbf{P}_2\mathbf{Q}_2\mathbf{z}$$

we can transform the coefficient matrix of the original state space equation into the Jordan canonical form. $\mathbf{T}$ is given by

$$\mathbf{T} = \lim_{\Delta\lambda_1 \to 0} \mathbf{P}_2\mathbf{Q}_2 = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_1 + a_1 & 1 & 0 \\ \lambda_1^2 + a_1\lambda_1 + a_2 & 2\lambda_1 + a_1 & 1 \end{bmatrix}$$

Problem A-3-12 Obtain a transformation matrix $\mathbf{T}$ which transforms the following state space equation into the Jordan canonical form:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 6 & 1 & 0 \\ -12 & 0 & 1 \\ 8 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Solution The characteristic equation becomes

$$\lambda^3 - 6\lambda^2 + 12\lambda - 8 = (\lambda - 2)^3 = (\lambda - \lambda_1)^3 = 0$$

Hence, from Eq. (3–53) $\mathbf{T}$ is given by

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_1 + a_1 & 1 & 0 \\ \lambda_1^2 + a_1\lambda_1 + a_2 & 2\lambda_1 + a_1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -4 & 1 & 0 \\ 4 & -2 & 1 \end{bmatrix}$$

PROBLEMS

Problem B-3-1 Assume that $\mathbf{A}$ is an $n \times n$ matrix and the eigenvalues of $\mathbf{A}$ involve k pairs of complex ones, namely

$$\lambda_1, \bar{\lambda}_1, \lambda_2, \bar{\lambda}_2, \ldots, \lambda_k, \bar{\lambda}_k, \lambda_{2k+1}, \lambda_{2k+2}, \ldots, \lambda_n$$

where the first $2k$ eigenvalues are the k conjugate complex pairs and the last $n - 2k$ eigenvalues are real. Assume also that they are all distinct.

Show that it is possible to transform $\mathbf{A}$ into a diagonal matrix, or

$$\mathbf{P}_1^{-1}\mathbf{A}\mathbf{P}_1 = \text{diag}\,(\lambda_1, \bar{\lambda}_1, \ldots, \lambda_k, \bar{\lambda}_k, \lambda_{2k+1}, \ldots, \lambda_n)$$

The columns corresponding to λ_i and $\bar{\lambda}_i$ of the transformation matrix $\mathbb{P}_1$ consist of conjugate elements. Show also that by a second transformation $\mathbb{P}_2$ it is possible to transform $\mathbf{A}_1 = \mathbf{P}_1^{-1}\mathbf{A}\mathbf{P}_1$ into the following:

$$\mathbf{P}_2^{-1}\mathbf{A}_1\mathbf{P}_2 = \mathbf{A}_2 = \begin{bmatrix} G_1 & & & & & 0 \\ & G_2 & & & & \\ & & \ddots & & & \\ & & & G_k & & \\ & & & & \lambda_{2k+1} & \\ & & & & & \ddots \\ 0 & & & & & \lambda_n \end{bmatrix}$$

where

$$\mathbf{G}_i = \begin{bmatrix} \sigma_i & \omega_i \\ -\omega_i & \sigma_i \end{bmatrix}$$

$$\lambda_i = \sigma_i + j\omega_i$$

This means that by the transformation $\mathbf{P} = \mathbf{P}_1\mathbf{P}_2$, $\mathbf{A}$ is transformed into $\mathbf{A}_2$, or

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{A}_2$$

Show that the transformation matrix $\mathbf{P}$ is a real matrix.

Problem B-3-2 Diagonalize the coefficient matrix of the following state space equation:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 4 & 2 & 1 \\ 7 & 3 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Problem B-3-3 Prove that if a 3×3 real symmetric matrix $\mathbf{A}$ where

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \qquad (a_{ij} = a_{ji})$$

possesses eigenvalues $\lambda_1, \lambda_1, \lambda_3$, then the elements a_{ij} will satisfy the following relationships:

$$\frac{a_{11} - \lambda_1}{a_{12}} = \frac{a_{12}}{a_{22} - \lambda_1} = \frac{a_{13}}{a_{23}}$$

$$\frac{a_{11} - \lambda_1}{a_{13}} = \frac{a_{12}}{a_{23}} = \frac{a_{13}}{a_{33} - \lambda_1}$$

Problem B-3-4 Find a matrix $\mathbf{S}$ such that

$$\mathbf{SA} = \mathbf{JS}$$

where

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 3 \\ 4 & 1 & 2 \\ -\frac{24}{13} & \frac{10}{13} & 2 \end{bmatrix}, \qquad \mathbf{J} = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Find also a matrix $\mathbf{T}$ such that

$$\mathbf{TA} = \hat{\mathbf{J}}\mathbf{T}$$

where

$$\hat{\mathbf{J}} = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Problem B-3-5 Obtain a state space equation of a system defined by

$$\dddot{x} + 9\ddot{x} + 23\dot{x} + 15x = 0$$

Simplify the coefficient matrix of the state space equation into the diagonal form.

Problem B-3-6 Consider the following equation:

$$\overset{(n)}{x} + a_1 \overset{(n-1)}{x} + \cdots + a_{n-1}\dot{x} + a_n x = 0$$

By defining

$$\begin{aligned} y_1 &= x \\ y_2 &= \dot{x} + a_1 x \\ y_3 &= \ddot{x} + a_1 \dot{x} + a_2 x \\ &\cdots \qquad \cdots \\ y_n &= \overset{(n-1)}{x} + a_1 \overset{(n-2)}{x} + \cdots + a_{n-1} x \end{aligned}$$

the given differential equation becomes

$$\dot{\mathbf{y}} = \mathbf{Ay}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} -a_1 & 1 & 0 & \cdots & 0 \\ -a_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -a_{n-1} & 0 & 0 & \cdots & 1 \\ -a_n & 0 & 0 & \cdots & 0 \end{bmatrix}$$

Define

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_1 & 1 & 0 & \cdots & 0 \\ a_2 & a_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ a_{n-1} & a_{n-2} & a_{n-3} & \cdots & 1 \end{bmatrix}$$

Show that

$$\mathbf{B}^{-1}\mathbf{AB} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_1 \end{bmatrix}$$

REFERENCES

3–1 Ogata, K., Class Notes on Control Theory, Lecture, Yokohama National University, 1960–61.

3–2 Turnbull, H. W., and A. C. Aitken, *An Introduction to the Theory of Canonical Matrices*, London: Blackie and Son, Ltd., 1932.

3–3 Ichikawa, K., "Coordinate Transformation," *Tech. Reports Inst. Automatic Control* (Faculty of Engineering, Nagoya University), **10** (1962), 61–77.

3–4 Desoer, C. A., "An Introduction to State Space Techniques in Linear Systems," *JACC* Preprint (1962), paper No. 10-2.

3–5 Enns, M., J. R. Greenwood III, J. E. Matheson, and F. T. Thompson, "Practical Aspects of State Space Methods: Part 1—System Formulation and Reduction," *JACC* Preprint (1964) 494–500.

3–6 Parlett, B., "Applications of Laguerre's Method to the Matrix Eigenvalue Problem," *Tech. Report*, Applied Mathematics and Statistics Laboratories, Stanford University, Calif., **21**, (1962).

4

State Space Representation of Control Systems

4-1 INTRODUCTION

In state space analysis of systems we frequently wish to obtain a state space representation of the following scalar differential equation:

$$\overset{(n)}{x} + a_1\overset{(n-1)}{x} + \cdots + a_{n-1}\dot{x} + a_n x = b_0\overset{(n)}{u} + b_1\overset{(n-1)}{u} + \cdots + b_{n-1}\dot{u} + b_n u$$

where the coefficients a_i and b_j are constants, or conversely, we wish to transform state space equations of the form $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u$ into scalar differential equations.

In this chapter we shall be mainly concerned with deriving state space equations from scalar differential equations and vice versa. Specifically, in Section 4–2 we shall derive state space representation of systems described by nth-order scalar differential equations. Section 4–3 discusses a converse problem of obtaining scalar differential equations from state space equations. In Section 4–4 we shall present the partial-fraction expansion technique for obtaining state space equations of transfer functions. Finally, Section 4–5 deals with difference equations and their state space representation.

4-2 STATE SPACE REPRESENTATION OF SYSTEMS DESCRIBED BY SCALAR DIFFERENTIAL EQUATIONS

In this section we shall derive state space representation of systems described by scalar differential equations. We shall limit our discussions only to linear systems. Notice that if a state space equation and a scalar differential equation are to represent the same physical system, then these two equations must have the same characteristic equation. It is clear that if the scalar differential equation is of the order n, then the corresponding state space equation must consist of n first-order differential equations.

In deriving state space equations from scalar differential equations, we should neither divide the common terms in the left-hand side and right-hand side of the equation, nor cancel the common terms in the numerator and denominator of the transfer function. For example, consider the following equation which describes a forced dynamic system:

$$\dddot{x} + 3\ddot{x} + 2\dot{x} = \dot{u} + u$$

where x and u are the output and the input of the system, respectively. This differential equation corresponds to the transfer function

$$\frac{X(s)}{U(s)} = \frac{s+1}{s(s+1)(s+2)}$$

If the common terms $s + 1$ in the numerator and in the denominator are canceled, then the transfer function is reduced to

$$\frac{X(s)}{U(s)} = \frac{1}{s(s+2)}$$

which corresponds to

$$\ddot{x} + 2\dot{x} = u$$

The original differential equation has the characteristic equation of

$$\lambda^3 + 3\lambda^2 + 2\lambda = 0$$

whereas the simplified differential equation has the characteristic equation of

$$\lambda^2 + 2\lambda = 0$$

Hence the original third-order differential equation and the simplified second-order differential equation do not represent the same system.

From the foregoing analysis, it is clear that in deriving a state space equation of the system we should not divide the left-hand side and right-hand side of the system's scalar differential equation by the common factor. Keeping this in mind, we shall derive state space representation of the following scalar differential equation:

$$\overset{(n)}{x} + a_1\overset{(n-1)}{x} + \cdots + a_{n-1}\dot{x} + a_n x = b_0\overset{(n)}{u} + b_1\overset{(n-1)}{u} + \cdots + b_{n-1}\dot{u} + b_n u \quad (4\text{–}1)$$

where the a_i and the b_j are constants. Note that there is no loss in generality to assume the coefficient of the highest-order derivative of x to be unity. Any of the coefficients a_i and b_j may be zero. We shall first consider a simple case where $b_0 = b_1 = \cdots = b_{n-1} = 0$ and $b_n = b \neq 0$. Then we shall treat a general case of Eq. (4–1).

Nth-order scalar differential equation with simple forcing term

Consider the following equation:

$$\overset{(n)}{x} + a_1 \overset{(n-1)}{x} + \cdots + a_{n-1}\dot{x} + a_n x = bu \tag{4–2}$$

By letting

$$\begin{aligned} x_1 &= x \\ x_2 &= \dot{x}_1 \\ &\cdots \\ x_n &= \overset{(n-1)}{x} \end{aligned}$$

Eq. (4–2) can be written as

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u \tag{4–3}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_{n-1} \\ x_n \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & 1 \\ -a_n & -a_{n-1} & \cdots & -a_1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix}$$

Transformation of $\mathbf{A}$ into the Jordan canonical form will ease computational problems and will provide the insight important to system compensation. In the following discussion it is assumed that the eigenvalues of the characteristic equation are real and distinct. (Extension to the case where the given equation involves complex conjugate eigenvalues is obvious.) The case of multiple eigenvalues will be treated later.

If $\mathbf{x}$ is transformed into $\mathbf{y}$ by the transformation matrix $\mathbf{P}$, or $\mathbf{x} = \mathbf{P}\mathbf{y}$, then the state space equation becomes

$$\dot{\mathbf{x}} = \mathbf{P}\dot{\mathbf{y}} = \mathbf{A}\mathbf{P}\mathbf{y} + \mathbf{B}u$$

or

$$\begin{aligned} \dot{\mathbf{y}} &= \mathbf{P}^{-1}\mathbf{A}\mathbf{P}\mathbf{y} + \mathbf{P}^{-1}\mathbf{B}u \\ &= \mathbf{D}\mathbf{y} + \mathbf{P}^{-1}\mathbf{B}u \end{aligned} \tag{4–4}$$

where $\mathbf{D}$ is a diagonal matrix. The forcing term $\mathbf{B}u$ is thus transformed into $\mathbf{P}^{-1}\mathbf{B}u$. Since $\mathbf{P}^{-1}\mathbf{B}u$ can be obtained as

$$\mathbf{P}^{-1}\mathbf{B}u = \mathbf{P}^{-1}\begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ u \end{bmatrix} = \begin{bmatrix} \frac{\Delta_{n1}}{|\mathbf{P}|}u \\ \frac{\Delta_{n2}}{|\mathbf{P}|}u \\ \cdot \\ \cdot \\ \cdot \\ \frac{\Delta_{nn}}{|\mathbf{P}|}u \end{bmatrix}$$

where Δ_{nj} is the cofactor of the (n, j)th element of $\mathbf{P}$, Eq. (4–4) can be written in the following form:

$$\begin{aligned} \dot{y}_1 &= \lambda_1 y_1 + \frac{\Delta_{n1}}{|\mathbf{P}|}u \\ \dot{y}_2 &= \lambda_2 y_2 + \frac{\Delta_{n2}}{|\mathbf{P}|}u \\ &\cdots \quad \cdots \\ \dot{y}_n &= \lambda_n y_n + \frac{\Delta_{nn}}{|\mathbf{P}|}u \end{aligned} \tag{4–5}$$

Notice that in the presence of the forcing function the resulting transformed form is not unique in the sense that the coefficients of the forcing function terms depend on the choice of the transformation matrix $\mathbf{P}$. This implies that if we choose $\mathbf{P}$ properly, each coefficient of the forcing function u can be made unity. We shall see this in the following. If we write

$$\mathbf{P}^{-1}\mathbf{B} = \begin{bmatrix} v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_n \end{bmatrix} = \begin{bmatrix} v_1 & 0 & 0 & \cdots & 0 \\ 0 & v_2 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & v_n \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix} = \mathbf{V}\mathbf{F}$$

where

$$\mathbf{V} = \begin{bmatrix} v_1 & 0 & 0 & \cdots & 0 \\ 0 & v_2 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & v_n \end{bmatrix} = \text{nonsingular matrix} \tag{4–6}$$

$$\mathbf{F} = \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix}$$

then Eq. (4–4) can be modified into

$$\dot{\mathbf{y}} = \mathbf{D}\mathbf{y} + \mathbf{V}\mathbf{F}u$$

If a new variable $\mathbf{z}$ is defined by $\mathbf{y} = \mathbf{V}\mathbf{z}$, then

$$\dot{\mathbf{z}} = \mathbf{V}^{-1}\mathbf{D}\mathbf{V}\mathbf{z} + \mathbf{F}u \tag{4–7}$$

Since $\mathbf{V}$ is a diagonal matrix we have

$$\mathbf{V}^{-1}\mathbf{D}\mathbf{V} = \mathbf{D}$$

Equation (4–7) can therefore be simplified into

$$\dot{\mathbf{z}} = \mathbf{D}\mathbf{z} + \mathbf{F}u$$

namely,

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \cdot \\ \cdot \\ \cdot \\ \dot{z}_n \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \cdot \\ \cdot \\ \cdot \\ z_n \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix} [u] \tag{4–8}$$

or

$$\begin{aligned} \dot{z}_1 &= \lambda_1 z_1 + u \\ \dot{z}_2 &= \lambda_2 z_2 + u \\ &\cdot \quad \cdot \quad \cdot \quad \cdot \\ \dot{z}_n &= \lambda_n z_n + u \end{aligned} \tag{4–9}$$

Thus, it has been shown that by the transformations

$$\mathbf{x} = \mathbf{P}\mathbf{y}, \qquad \mathbf{y} = \mathbf{V}\mathbf{z}$$

Eq. (4–2) can be transformed into the form given by Eq. (4–8) or Eq. (4–9). The initial condition terms are transformed into

$$\mathbf{z}(0) = \mathbf{V}^{-1}\mathbf{P}^{-1}\mathbf{x}(0)$$

Notice that $\mathbf{z}(t)$ is uniquely determined by $\mathbf{x}(0)$ and $u(t)$.

We shall next consider the case where $\mathbf{A}$ involves multiple eigenvalues. In this case the form given by Eq. (4–8) must be modified. Consider the following system:

$$\dddot{x} - 3\lambda_1\ddot{x} + 3\lambda_1^2\dot{x} - \lambda_1^3 = u \tag{4–10}$$

The three roots of the characteristic equation are equal, or $\lambda_1 = \lambda_2 = \lambda_3$. In Chapter 3 we have shown that the simplest form we can get of the homogeneous equation of Eq. (4–10) with $\lambda_1 = \lambda_2 = \lambda_3$ is

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 1 & 0 \\ 0 & \lambda_1 & 1 \\ 0 & 0 & \lambda_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

For the nonhomogeneous equation we may write

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 1 & 0 \\ 0 & \lambda_1 & 1 \\ 0 & 0 & \lambda_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} a \\ b \\ c \end{bmatrix} [u]$$

where a, b, and c are constants to be determined. By eliminating x_2 and x_3 from this last equation and writing $x_1 = x$ we obtain

$$\begin{aligned} \dddot{x} - 3\lambda_1\ddot{x} + 3\lambda_1^2\dot{x} - \lambda_1^3 x \\ = a\ddot{u} + (b - 2a\lambda_1)\dot{u} + (\lambda_1^2 a - \lambda_1 b + c)u \end{aligned}$$

The right-hand side of this equation must be equal to u. Therefore,

$$a\ddot{u} + (b - 2a\lambda_1)\dot{u} + (\lambda_1^2 a - \lambda_1 b + c)u = u$$

which yields

$$a = b = 0, \qquad c = 1$$

Hence the simplest form of Eq. (4–10) can be written as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 1 & 0 \\ 0 & \lambda_1 & 1 \\ 0 & 0 & \lambda_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} [u]$$

Similarly, the simplest form of Eq. (4–3) when $\mathbf{A}$ involves, for example, seven eigenvalues $(\lambda_1, \lambda_1, \lambda_1, \lambda_4, \lambda_4, \lambda_6, \lambda_7)$ is illustrated below:

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \\ \dot{y}_4 \\ \dot{y}_5 \\ \dot{y}_6 \\ \dot{y}_7 \end{bmatrix} = \left[\begin{array}{ccc:ccc:cc} \lambda_1 & 1 & 0 & & & & 0 \\ 0 & \lambda_1 & 1 & & & & \\ 0 & 0 & \lambda_1 & & & & \\ \hdashline & & & \lambda_4 & 1 & & \\ & & & 0 & \lambda_4 & & \\ \hdashline & & & & & \lambda_6 & 0 \\ 0 & & & & & 0 & \lambda_7 \end{array}\right] \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} [u]$$

(For the case of multiple roots, refer also to Section 4–4.)

Example 4-1 Consider the system described by the following scalar differential equation:

$$\dddot{x} + 6\ddot{x} + 11\dot{x} + 6x = u \tag{4–11}$$

Obtain the state space representation of the system equation in which the coefficient matrix of the state vector is in a diagonal form and the elements of the coefficient matrix of the forcing function u are unity.

Let $x_1 = x$, $x_2 = \dot{x}$, and $x_3 = \ddot{x}$. Then

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & -11 & -6 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

The eigenvalues of **A** are -1, -2, and -3. Let y_1, y_2, y_3 be the principal coordinates. By means of the following transformation

$$\mathbf{x} = \mathbf{P}\mathbf{y}$$

where

$$\mathbf{P} = \begin{bmatrix} 1 & 1 & 1 \\ \lambda_1 & \lambda_2 & \lambda_3 \\ \lambda_1^2 & \lambda_2^2 & \lambda_3^2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ -1 & -2 & -3 \\ 1 & 4 & 9 \end{bmatrix} \tag{4-12}$$

we obtain

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + \mathbf{P}^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} [u]$$

or

$$\dot{\mathbf{y}} = \mathbf{D}\mathbf{y} + \mathbf{P}^{-1}\mathbf{B}u$$

Once the roots of the characteristic equation are known it is not necessary to evaluate $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$, since $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ is a diagonal matrix with only eigenvalues appearing on the diagonal. It is necessary, however, to find $\mathbf{P}^{-1}$ since the forcing function term is modified by $\mathbf{P}^{-1}$. $\mathbf{P}^{-1}$ is obtained as

$$\mathbf{P}^{-1} = \begin{bmatrix} 3 & \frac{5}{2} & \frac{1}{2} \\ -3 & -4 & -1 \\ 1 & \frac{3}{2} & \frac{1}{2} \end{bmatrix}$$

and

$$\mathbf{P}^{-1}\mathbf{B}u = \begin{bmatrix} \frac{1}{2} \\ -1 \\ \frac{1}{2} \end{bmatrix} [u]$$

Equation (4–11) is thus transformed into

$$\begin{aligned} \dot{y}_1 &= -y_1 + \tfrac{1}{2}u \\ \dot{y}_2 &= -2y_2 - u \\ \dot{y}_3 &= -3y_3 + \tfrac{1}{2}u \end{aligned} \tag{4-13}$$

To transform Eq..(4–13) into the form of Eq. (4–9), let

$$\mathbf{y} = \mathbf{V}\mathbf{z}$$

or

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

Since

$$\mathbf{V}^{-1} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

and

$$\mathbf{V}^{-1}\mathbf{P}^{-1}\mathbf{B}u = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} \frac{1}{2} \\ -1 \\ \frac{1}{2} \end{bmatrix} [u] = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} [u]$$

Eq. (4–13) can be transformed into the following form:

$$\begin{aligned} \dot{z}_1 &= -z_1 + u \\ \dot{z}_2 &= -2z_2 + u \\ \dot{z}_3 &= -3z_3 + u \end{aligned} \tag{4-14}$$

or

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} [u]$$

The initial condition $\mathbf{z}(0)$ is given by

$$\mathbf{z}(0) = \mathbf{V}^{-1}\mathbf{P}^{-1}\mathbf{x}(0)$$

For a given u the solution to Eq. (4–14) can be obtained easily. To obtain the solution in terms of $\mathbf{x}$, we must transform $\mathbf{z}$ into $\mathbf{x}$. In the present problem this can be done by the following transformation:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ -1 & -2 & -3 \\ 1 & 4 & 9 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

Nth-order scalar differential equation with forcing terms involving derivatives of forcing function u We shall next consider state space representation of a system described by Eq. (4–1). In such a system we should choose a set of variables such that in the state space the motion does not exhibit infinite jumps. Determination of a suitable state vector is possible, since the derivatives of x and u provide all the necessary information for the determination of the future trajectory of such a system. The components of each possible state vector consist of suitable combinations of derivatives of various orders of x and u. Several methods are available for

finding possible state vectors for such a case. We shall discuss three methods in this section and another method in Section 4–4.

Method 1 We shall demonstrate the first method by use of a third-order system. Then, we shall extend the method to the nth-order system.

Consider the following equation:

$$\dddot{x} + a_1\ddot{x} + a_2\dot{x} + a_3x = b_0\ddot{u} + b_1\dot{u} + b_2\dot{u} + b_3u$$

In terms of the operator $p = d/dt$, this equation can be written as

$$(p^3 + a_1p^2 + a_2p + a_3)x = (b_0p^3 + b_1p^2 + b_2p + b_3)u$$

or

$$p^3(x - b_0u) + p^2(a_1x - b_1u) + p(a_2x - b_2u) + (a_3x - b_3u) = 0$$

from which

$$x = b_0u + \frac{1}{p}(b_1u - a_1x) + \frac{1}{p^2}(b_2u - a_2x) + \frac{1}{p^3}(b_3u - a_3x)$$

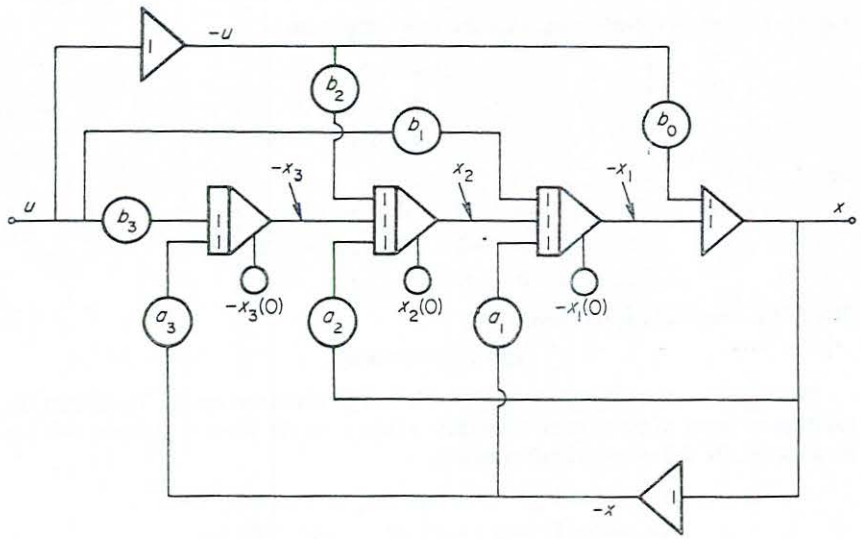


Figure 4-1 Analog computer diagram for third-order differential equation $\dddot{x} + a_1\ddot{x} + a_2\dot{x} + a_3x = b_0\dddot{u} + b_1\ddot{u} + b_2\dot{u} + b_3u$.

An analog computer diagram for this last equation is shown in Fig. 4–1. The output signals from the integrators constitute a set of state variables. Let us define the output signals from the integrators as $-x_1$, x_2, and $-x_3$ as shown in Fig. 4–1. Then x_1, x_2, and x_3 are a set of state variables for the present system. In terms of the original variables x and u, the state variables x_1, x_2, and x_3 are given by

$$\begin{aligned} x_1 &= x - b_0u \\ x_2 &= p(x - b_0u) + (a_1x - b_1u) \\ x_3 &= p^2(x - b_0u) + p(a_1x - b_1u) + (a_2x - b_2u) \end{aligned} \tag{4–15}$$

Note that

$$px_3 = -(a_3x - b_3u) \tag{4–16}$$

In terms of matrix representation, the state vector $\mathbf{x}$ where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

can be written as

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ a_1 & 1 & 0 \\ a_2 & a_1 & 1 \end{bmatrix} \begin{bmatrix} x \\ \dot{x} \\ \ddot{x} \end{bmatrix} + \begin{bmatrix} -b_0 & 0 & 0 \\ -b_1 & -b_0 & 0 \\ -b_2 & -b_1 & -b_0 \end{bmatrix} \begin{bmatrix} u \\ \dot{u} \\ \ddot{u} \end{bmatrix}$$

Equations (4–15) and Eq. (4–16) may be modified into

$$\begin{aligned} x_1 &= x - b_0u \\ x_2 &= \dot{x}_1 + a_1x - b_1u \\ x_3 &= \dot{x}_2 + a_2x - b_2u \\ \dot{x}_3 &= -a_3x + b_3u \end{aligned} \tag{4–17}$$

Eliminating x from Eq. (4–17) by use of the relationship $x = x_1 + b_0u$ we obtain

$$\begin{aligned} \dot{x}_1 &= -a_1x_1 + x_2 + (b_1 - a_1b_0)u \\ \dot{x}_2 &= -a_2x_1 + x_3 + (b_2 - a_2b_0)u \\ \dot{x}_3 &= -a_3x_1 + (b_3 - a_3b_0)u \end{aligned}$$

or

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} -a_1 & 1 & 0 \\ -a_2 & 0 & 1 \\ -a_3 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} b_1 - a_1b_0 \\ b_2 - a_2b_0 \\ b_3 - a_3b_0 \end{bmatrix} [u] \tag{4–18}$$

where

$$x_1 = x - b_0u$$

(Notice that although the state vector $\mathbf{x}$ may be continuous, the original variable x may have discontinuities at points where the value of u jumps.)

Equation (4–18) may be written as

$$\dot{\mathbf{x}} = \mathbf{C}\mathbf{x} + \mathbf{H}u \tag{4–19}$$

By use of an appropriate transformation

$$\mathbf{x} = \mathbf{P}\mathbf{y}$$

Eq. (4–19) can be modified into

$$\dot{\mathbf{y}} = \mathbf{P}^{-1}\mathbf{C}\mathbf{P}\mathbf{y} + \mathbf{P}^{-1}\mathbf{H}u$$

where $\mathbf{P}^{-1}\mathbf{C}\mathbf{P}$ is in the Jordan canonical form; in particular if the eigenvalues of the characteristic equation are distinct, it is in the diagonal form. If necessary, the coefficient matrix of u can be normalized by use of an additional transformation

$$\mathbf{y} = \mathbf{V}\mathbf{z}$$

where $\mathbf{V}$ is defined by Eq. (4–6). Then the simplified equation of the distinct eigenvalue system is

$$\dot{\mathbf{z}} = \mathbf{D}\mathbf{z} + \mathbf{F}u \tag{4–20}$$

where

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix}, \qquad \mathbf{F} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

The initial condition terms are transformed into

$$\mathbf{z}(0) = \mathbf{V}^{-1}\mathbf{P}^{-1}\mathbf{x}(0) \tag{4–21}$$

where $\mathbf{x}(0)$ is given by

$$\mathbf{x}(0) = \begin{bmatrix} x_1(0) \\ x_2(0) \\ x_3(0) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ a_1 & 1 & 0 \\ a_2 & a_1 & 1 \end{bmatrix} \begin{bmatrix} x(0) \\ \dot{x}(0) \\ \ddot{x}(0) \end{bmatrix} + \begin{bmatrix} -b_0 & 0 & 0 \\ -b_1 & -b_0 & 0 \\ -b_2 & -b_1 & -b_0 \end{bmatrix} \begin{bmatrix} u(0) \\ \dot{u}(0) \\ \ddot{u}(0) \end{bmatrix}$$

Example 4-2 Given the system equation

$$\dddot{x} + 6\ddot{x} + 11\dot{x} + 6x = \dddot{u} + 8\ddot{u} + 17\dot{u} + 8u \tag{4–22}$$

obtain the state space representation in the form of Eq. (4–20).

By use of the transformation defined by Eq. (4–15), Eq. (4–22) becomes as follows:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} -6 & 1 & 0 \\ -11 & 0 & 1 \\ -6 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 2 \\ 6 \\ 2 \end{bmatrix} [u] \tag{4–23}$$

The roots of the characteristic equation are -1, -2, and -3. Let us transform Eq. (4–23) into an equation in $\mathbf{y}$ by the following transformation:

$$\mathbf{x} = \mathbf{P}\mathbf{y}$$

where $\mathbf{P}$ is defined by matrix (3–21) and given by

$$\mathbf{P} = \begin{bmatrix} 1 & 1 & 1 \\ 5 & 4 & 3 \\ 6 & 3 & 2 \end{bmatrix}$$

$\mathbf{P}^{-1}$ is then obtained as

$$\mathbf{P}^{-1} = -\tfrac{1}{2}\begin{bmatrix} -1 & 1 & -1 \\ 8 & -4 & 2 \\ -9 & 3 & -1 \end{bmatrix}$$

Hence

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ -4 & 2 & -1 \\ \frac{9}{2} & -\frac{3}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 2 \\ 6 \\ 2 \end{bmatrix} [u]$$

$$= \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix} [u] \tag{4–24}$$

By a further transformation

$$\mathbf{y} = \mathbf{V}\mathbf{z}$$

where

$$\mathbf{V} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Eq. (4–24) can be simplified, giving

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} [u]$$

The initial condition $\mathbf{z}(0)$ is defined by Eq. (4–21).

For a given u, the response $\mathbf{z}$ can be computed easily. Once $\mathbf{z}$ is found, then the response x in the original variable can be obtained immediately. Since

$$\mathbf{x} = \mathbf{P}\mathbf{V}\mathbf{z}$$

$$= \begin{bmatrix} 1 & 1 & 1 \\ 5 & 4 & 3 \\ 6 & 3 & 2 \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} -1 & 2 & 1 \\ -5 & 8 & 3 \\ -6 & 6 & 2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

and

$$x = x_1 + b_0 u$$

we obtain x in terms of z_1, z_2, z_3, and u as follows:

$$x = -z_1 + 2z_2 + z_3 + u$$

We shall postpone general discussions of system responses, or solutions of state space equations to Chapter 6.

Let us next consider a general case where the system differential equation is given by the following equation:

$$\overset{(n)}{x} + a_1\overset{(n-1)}{x} + \cdots + a_{n-1}\dot{x} + a_n x = b_0\overset{(n)}{u} + b_1\overset{(n-1)}{u} + \cdots + b_{n-1}\dot{u} + b_n u \tag{4–25}$$

where the coefficients a_i and b_j are constants. Let us define a set of state variables $x_1, x_2, \ldots, x_n$ in the following way:

$$\begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ a_1 & 1 & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n-1} & a_{n-2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} x \\ \dot{x} \\ \cdot \\ \cdot \\ \cdot \\ \overset{(n-1)}{x} \end{bmatrix} + \begin{bmatrix} -b_0 & 0 & \cdots & 0 \\ -b_1 & -b_0 & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ -b_{n-1} & -b_{n-2} & \cdots & -b_0 \end{bmatrix} \begin{bmatrix} u \\ \dot{u} \\ \cdot \\ \cdot \\ \cdot \\ \overset{(n-1)}{u} \end{bmatrix} \tag{4–26}$$

which is equivalent to

$$\begin{aligned} x_1 &= x - b_0 u \\ x_2 &= \dot{x}_1 + a_1 x - b_1 u \\ x_3 &= \dot{x}_2 + a_2 x - b_2 u \\ &\cdots \quad \cdots \\ x_n &= \dot{x}_{n-1} + a_{n-1} x - b_{n-1} u \\ \dot{x}_n &= -a_n x + b_n u \end{aligned} \tag{4–27}$$

Eliminating x from Eq. (4–27) by use of the relationship

$$x = x_1 + b_0 u \tag{4–28}$$

we obtain the following equation:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \cdot \\ \cdot \\ \cdot \\ \dot{x}_{n-1} \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} -a_1 & 1 & 0 & \cdots & 0 \\ -a_2 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ -a_{n-1} & 0 & 0 & \cdots & 1 \\ -a_n & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_{n-1} \\ x_n \end{bmatrix} + \begin{bmatrix} b_1 - a_1 b_0 \\ b_2 - a_2 b_0 \\ \cdot \\ \cdot \\ \cdot \\ b_{n-1} - a_{n-1} b_0 \\ b_n - a_n b_0 \end{bmatrix} [u] \tag{4–29}$$

where x_1 and x, the original variable of the equation, are related by Eq. (4–28).

The initial condition $\mathbf{x}(0)$ can be obtained in terms of $x(0), \dot{x}(0), \ldots, \overset{(n-1)}{x}(0)$ and $u(0), \dot{u}(0), \ldots, \overset{(n-1)}{u}(0)$ by use of Eq. (4–26).

If the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of the $n \times n$ matrix in Eq. (4–29) are distinct, the system equation can be simplified by the use of the following transformation matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 + a_1 & \lambda_2 + a_1 & \cdots & \lambda_n + a_1 \\ \lambda_1^2 + a_1\lambda_1 + a_2 & \lambda_2^2 + a_1\lambda_2 + a_2 & \cdots & \lambda_n^2 + a_1\lambda_n + a_2 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \lambda_1^{n-1} + a_1\lambda_1^{n-2} + \cdots + a_{n-1} & \lambda_2^{n-1} + a_1\lambda_2^{n-2} + \cdots + a_{n-1} & \cdots & \lambda_n^{n-1} + a_1\lambda_n^{n-2} + \cdots + a_{n-1} \end{bmatrix}$$

(For the derivation of this transformation matrix, refer to Prob. A-3-10.) As an example, consider the system in Example 4–2. For this system since $\lambda_1 = -1, \lambda_2 = -2, \lambda_3 = -3$, and $a_1 = 6, a_2 = 11, a_3 = 6$, we obtain

$$\mathbf{P} = \begin{bmatrix} 1 & 1 & 1 \\ \lambda_1 + a_1 & \lambda_2 + a_1 & \lambda_3 + a_1 \\ \lambda_1^2 + a_1\lambda_1 + a_2 & \lambda_2^2 + a_1\lambda_2 + a_2 & \lambda_3^2 + a_1\lambda_3 + a_2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 1 & 1 \\ 5 & 4 & 3 \\ 6 & 3 & 2 \end{bmatrix}$$

This is the same $\mathbf{P}$ matrix as obtained in Example 4–2. If the eigenvalues of the $n \times n$ matrix in Eq. (4–29) involve multiple eigenvalues λ_1 of multiplicity m, then the transformation matrix becomes as follows: The elements in the jth column $(2 \leqslant j \leqslant m)$ is $(j-1)$th derivative (with respect to λ_1) of the corresponding elements in the first column divided by $(j-1)!$. For example, if the first three eigenvalues are equal and other eigenvalues are different from each other, then $\mathbf{P}$ becomes as follows:

$$
\mathbf{P} = \begin{bmatrix}
1 & 0 & 0 & 1 & \cdots \\
\lambda_1 + a_1 & 1 & 0 & \lambda_4 + a_1 & \cdots \\
\lambda_1^2 + a_1\lambda_1 + a_2 & 2\lambda_1 + a_1 & 1 & \lambda_4^2 + a_1\lambda_4 + a_2 & \cdots \\
g_3(\lambda_1) & 3\lambda_1^2 + 2a_1\lambda_1 + a_2 & 3\lambda_1 + a_1 & g_3(\lambda_4) & \cdots \\
\vdots & \vdots & \vdots & \vdots & \\
g_{n-2}(\lambda_1) & \dfrac{d}{d\lambda_1} g_{n-2}(\lambda_1) & \dfrac{1}{2!}\dfrac{d^2}{d\lambda_1^2} g_{n-2}(\lambda_1) & g_{n-2}(\lambda_4) & \cdots \\
g_{n-1}(\lambda_1) & \dfrac{d}{d\lambda_1} g_{n-1}(\lambda_1) & \dfrac{1}{2!}\dfrac{d^2}{d\lambda_1^2} g_{n-1}(\lambda_1) & g_{n-1}(\lambda_4) & \cdots
\end{bmatrix}
\begin{bmatrix}
1 \\
\lambda_n + a_1 \\
\lambda_n^2 + a_1\lambda_n + a_2 \\
g_3(\lambda_n) \\
\vdots \\
g_{n-2}(\lambda_n) \\
g_{n-1}(\lambda_n)
\end{bmatrix}
$$

where

$$g_{n-1}(\lambda) = \lambda^{n-1} + a_1\lambda^{n-2} + \cdots + a_{n-2}\lambda + a_{n-1}$$

(For the case of multiple eigenvalue systems refer to Prob. A-3-11.)

Method 2 We shall begin by demonstrating the second method by use of a third-order system. Consider the following equation:

$$\dddot{x} + a_1\ddot{x} + a_2\dot{x} + a_3 x = b_0\ddot{u} + b_1\ddot{u} + b_2\dot{u} + b_3 u \tag{4–30}$$

We wish to transform Eq. (4–30) into the following form:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_3 & -a_2 & -a_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} [u] \tag{4–31}$$

This is possible if we define

$$\begin{aligned} \dot{x}_1 &= x_2 + c_1 u \\ \dot{x}_2 &= x_3 + c_2 u \\ \dot{x}_3 &= -a_3 x_1 - a_2 x_2 - a_1 x_3 + c_3 u \end{aligned} \tag{4–32}$$

and

$$x = x_1 + c_0 u \tag{4–33}$$

The problem then becomes that of determining constants c_0, c_1, c_2, and c_3 in terms of the a's and b's. From Eq. (4–32) and Eq. (4–33), we obtain for $\dot{x}$

$$\dot{x} = x_2 + c_0\dot{u} + c_1 u \tag{4–34}$$

Similarly for $\ddot{x}$

$$\ddot{x} = x_3 + c_0\ddot{u} + c_1\dot{u} + c_2 u \tag{4–35}$$

and for $\dddot{x}$

$$\dddot{x} = -a_3 x_1 - a_2 x_2 - a_1 x_3 + c_0\dddot{u} + c_1\ddot{u} + c_2\dot{u} + c_3 u \tag{4–36}$$

Multiplying Eq. (4–35) by a_1, Eq. (4–34) by a_2, and Eq. (4–33) by a_3 and then adding these to Eq. (4–36), we obtain

$$\begin{aligned} \dddot{x} + a_1\ddot{x} + a_2\dot{x} + a_3 x &= -a_3 x_1 - a_2 x_2 - a_1 x_3 + c_0\dddot{u} + c_1\ddot{u} + c_2\dot{u} + c_3 u \\ &\quad + a_1 x_3 + a_1 c_0\ddot{u} + a_1 c_1\dot{u} + a_1 c_2 u \\ &\quad + a_2 x_2 + a_2 c_0\dot{u} + a_2 c_1 u + a_3 x_1 + a_3 c_0 u \\ &= c_0\dddot{u} + (c_1 + a_1 c_0)\ddot{u} + (c_2 + a_1 c_1 + a_2 c_0)\dot{u} \\ &\quad + (c_3 + a_1 c_2 + a_2 c_1 + a_3 c_0)u \\ &= b_0\dddot{u} + b_1\ddot{u} + b_2\dot{u} + b_3 u \end{aligned}$$

Hence

$$\begin{aligned} (c_0 - b_0)\dddot{u} + (c_1 + a_1 c_0 - b_1)\ddot{u} + (c_2 + a_1 c_1 + a_2 c_0 - b_2)\dot{u} \\ + (c_3 + a_1 c_2 + a_2 c_1 + a_3 c_0 - b_3)u = 0 \end{aligned}$$

In order to determine c_0, c_1, c_2, and c_3, let us equate coefficients of $\dddot{u}$, $\ddot{u}$, $\dot{u}$, and u of this equation to zero, respectively. Then

$$\begin{aligned} c_0 - b_0 &= 0 \\ c_1 + a_1 c_0 - b_1 &= 0 \\ c_2 + a_1 c_1 + a_2 c_0 - b_2 &= 0 \\ c_3 + a_1 c_2 + a_2 c_1 + a_3 c_0 - b_3 &= 0 \end{aligned}$$

it follows that

$$\begin{aligned} c_0 &= b_0 \\ c_1 &= b_1 - a_1 b_0 \\ c_2 &= b_2 - a_2 b_0 - a_1(b_1 - a_1 b_0) \\ c_3 &= b_3 - a_3 b_0 - (a_2 - a_1^2)(b_1 - a_1 b_0) - a_1(b_2 - a_2 b_0) \end{aligned} \tag{4–37}$$

Hence we see that Eq. (4–30) can be transformed into Eq. (4–31) by means of the transformation given by Eq. (4–32), where c_1, c_2, and c_3 are defined by Eq. (4–37). The initial conditions are

$$x_1(0) = x(0) - c_0u(0)$$
$$x_2(0) = \dot{x}(0) - c_0\dot{u}(0) - c_1u(0) \tag{4-38}$$
$$x_3(0) = \ddot{x}(0) - c_0\ddot{u}(0) - c_1\dot{u}(0) - c_2u(0)$$

Note that a set of state variables obtainable by Method 1 and that obtainable by Method 2 are different. In the case of the third-order system the set of state variables obtained by Method 2 can be written as follows:

$$x_1 = x - c_0u$$
$$x_2 = \dot{x} - c_0\dot{u} - c_1u$$
$$x_3 = \ddot{x} - c_0\ddot{u} - c_1\dot{u} - c_2u$$

In terms of the coefficients a_i and b_j, we have

$$x_1 = x - b_0u$$
$$x_2 = p(x - b_0u) + (a_1b_0 - b_1)u \tag{4-39}$$
$$x_3 = p^2(x - b_0u) + p(a_1b_0 - b_1)u - a_1(a_1b_0 - b_1)u + (a_2b_0 - b_2)u$$

where $p = d/dt$. [Compare Eq. (4–39) with Eq. (4–15). The two sets of equations are different from each other. As a matter of course, however, both sets of variables x_1, x_2, and x_3 are valid sets of state variables for the system considered.]

Example 4-3 Given the following system:

$$\dddot{x} + 6\ddot{x} + 11\dot{x} + 6x = \dddot{u} + 8\ddot{u} + 17\dot{u} + 8u$$

obtain the state space equation in the form of Eq. (4–20).

By use of Eq. (4–39) the given equation can be transformed into the following form:

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix} = \begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ -6 & -11 & -6\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix} + \begin{bmatrix}2\\ -6\\ 16\end{bmatrix}[u] \tag{4-40}$$

By means of the transformation

$$\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix} = \begin{bmatrix}1 & 1 & 1\\ -1 & -2 & -3\\ 1 & 4 & 9\end{bmatrix}\begin{bmatrix}-1 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}z_1\\ z_2\\ z_3\end{bmatrix} \tag{4-41}$$

Eq. (4–40) can be transformed into

$$\begin{bmatrix}\dot{z}_1\\ \dot{z}_2\\ \dot{z}_3\end{bmatrix} = \begin{bmatrix}-1 & 0 & 0\\ 0 & -2 & 0\\ 0 & 0 & -3\end{bmatrix}\begin{bmatrix}z_1\\ z_2\\ z_3\end{bmatrix} + \begin{bmatrix}1\\ 1\\ 1\end{bmatrix}[u]$$

The initial condition z(0) can be obtained from Eq. (4–37), Eq. (4–38), and Eq. (4–41). The solution x to the original differential equation is then obtained as

$$x = x_1 + b_0u = -z_1 + 2z_2 + z_3 + u$$

Example 4-4 Let us obtain a state space equation of the following system:

$$(p + 1)(p + 2)(p + 3)x = (p + 1)u, \qquad p = \frac{d}{dt} \tag{4-42}$$

By letting

$$x = x_1$$
$$\dot{x}_1 = x_2 + c_1u$$
$$\dot{x}_2 = x_3 + c_2u$$
$$\dot{x}_3 = -6x_1 - 11x_2 - 6x_3 + c_3u$$

the given differential equation can be written as follows:

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix} = \begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ -6 & -11 & -6\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix} + \begin{bmatrix}c_1\\ c_2\\ c_3\end{bmatrix}[u]$$

Constants c_1, c_2, and c_3 are obtained as

$$c_1 = 0, \qquad c_2 = 1, \qquad c_3 = -5$$

Hence we obtain

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix} = \begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ -6 & -11 & -6\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix} + \begin{bmatrix}0\\ 1\\ -5\end{bmatrix}[u]$$

which can be simplified into

$$\begin{bmatrix}\dot{y}_1\\ \dot{y}_2\\ \dot{y}_3\end{bmatrix} = \begin{bmatrix}-1 & 0 & 0\\ 0 & -2 & 0\\ 0 & 0 & -3\end{bmatrix}\begin{bmatrix}y_1\\ y_2\\ y_3\end{bmatrix} + \begin{bmatrix}0\\ 1\\ 1\end{bmatrix}[u] \tag{4-43}$$

where **x** and **y** are related by

$$\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix} = \begin{bmatrix}1 & 1 & 1\\ -1 & -2 & -3\\ 1 & 4 & 9\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & -1\end{bmatrix}\begin{bmatrix}y_1\\ y_2\\ y_3\end{bmatrix}$$

Equation (4–43) indicates that the motion of y_1 does not depend on u. This may be clear, since in Eq. (4–42) the factor $(p + 1)$ appears on both sides of the equation. (Notice that if we cancel the terms $(p + 1)$ on both sides of Eq. (4–42) then the order of the equation becomes second which is different from the order of the original system. Since the original equation is given as a third-order equation, the corresponding state space equation must consist of three first-order differential equations.)

We can easily extend the present analysis to a general case of the nth-order system:

$$x^{(n)} + a_1x^{(n-1)} + \cdots + a_{n-1}\dot{x} + a_nx = b_0u^{(n)} + b_1u^{(n-1)} + \cdots + b_{n-1}\dot{u} + b_nu \tag{4-44}$$

By defining the following set of equations:

$$\begin{aligned}
\dot{x}_1 &= x_2 + c_1 u \\
\dot{x}_2 &= x_3 + c_2 u \\
&\cdots \\
\dot{x}_{n-1} &= x_n + c_{n-1} u \\
\dot{x}_n &= -a_n x_1 - a_{n-1} x_2 - \cdots - a_1 x_n + c_n u
\end{aligned}$$

and

$$x = x_1 + c_0 u$$

we can transform Eq. (4–44) into the following form:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_{n-1} \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} + \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_{n-1} \\ c_n \end{bmatrix} [u] \tag{4–45}$$

with initial conditions

$$\begin{aligned}
x_1(0) &= x(0) - c_0 u(0) \\
x_2(0) &= \dot{x}(0) - c_0 \dot{u}(0) - c_1 u(0) \\
x_3(0) &= \ddot{x}(0) - c_0 \ddot{u}(0) - c_1 \dot{u}(0) - c_2 u(0) \\
&\cdots \\
x_n(0) &= x^{(n-1)}(0) - c_0 u^{(n-1)}(0) - \cdots - c_{n-2} \dot{u}(0) - c_{n-1} u(0)
\end{aligned}$$

where $c_0, c_1, c_2, \ldots, c_n$ are determined from

$$\begin{aligned}
c_0 &= b_0 \\
c_1 + a_1 c_0 &= b_1 \\
c_2 + a_1 c_1 + a_2 c_0 &= b_2 \\
c_3 + a_1 c_2 + a_2 c_1 + a_3 c_0 &= b_3 \\
&\cdots \\
c_n + a_1 c_{n-1} + \cdots + a_{n-1} c_1 + a_n c_0 &= b_n
\end{aligned} \tag{4–46}$$

Equation (4–45) can be simplified, if necessary, by the straightforward technique discussed in this section.

Method 3 We shall demonstrate the third method again by use of a third-order system. The result can be easily extended to the nth-order system.

Consider the following equation:

$$\dddot{x} + a_1 \ddot{x} + a_2 \dot{x} + a_3 x = b_0 \ddot{u} + b_1 \dot{u} + b_2 \dot{u} + b_3 u \tag{4–47}$$

We shall show that one possible state space representation of this system is given by

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_3 & -a_2 & -a_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} [u] \tag{4–48}$$

where the original dependent variable x is given by

$$x = (b_3 - a_3 b_0)x_1 + (b_2 - a_2 b_0)x_2 + (b_1 - a_1 b_0)x_3 + b_0 u$$

To demonstrate this possibility we shall consider an analog computer simulation of the original system given by Eq. (4–47). Suppose we define

$$\begin{aligned}
u &= \dddot{y} + a_1 \ddot{y} + a_2 \dot{y} + a_3 y \\
x &= b_0 \dddot{y} + b_1 \ddot{y} + b_2 \dot{y} + b_3 y
\end{aligned} \tag{4–49}$$

Then elimination of y and its derivatives from Eq. (4–49) results in Eq. (4–47). Figure 4–2 shows realization of Eq. (4–47) or Eq. (4–49). Let us define

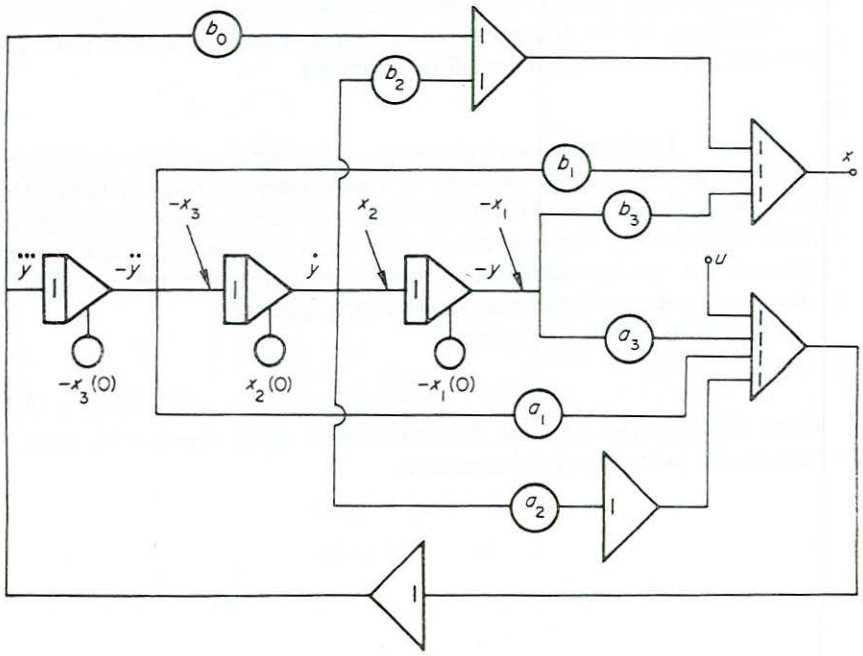


Figure 4-2 Analog computer diagram for third-order system defined by Eq. (4–47) or Eq. (4–49).

$$y = x_1$$
$$\dot{y} = x_2 \qquad (4\text{–}50)$$
$$\ddot{y} = x_3$$

Then from Eq. (4–49) and Eq. (4–50), we obtain

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = x_3$$
$$\dot{x}_3 = -a_3x_1 - a_2x_2 - a_1x_3 + u$$

and

$$x = (b_3 - a_3b_0)x_1 + (b_2 - a_2b_0)x_2 + (b_1 - a_1b_0)x_3 + b_0u$$

Hence Eq. (4–48) results. As seen from Eq. (4–48) essentially we have modified the original equation (4–47) into

$$\ddot{y} + a_1\ddot{y} + a_2\dot{y} + a_3y = u$$

The original variable x and the new variable y are related by

$$b_0\ddot{y} + b_1\ddot{y} + b_2\dot{y} + b_3y = x$$

Equation (4–48) can be simplified by techniques presented in this section.

Note that the analog computer diagram shown in Fig. 4–2 represents also a simulation of the following transfer function:

$$\frac{X(s)}{U(s)} = \frac{b_0s^3 + b_1s^2 + b_2s + b_3}{s^3 + a_1s^2 + a_2s + a_3}$$

provided that the initial conditions $x_1(0)$, $x_2(0)$, and $x_3(0)$ are set equal to zero.

Extension of the present method to the nth-order system is straightforward; therefore, details are left to the reader.

Linear time-varying system In presenting the preceding three methods we limited our systems to linear time-invariant systems. These methods, however, can be extended to treat linear time-varying systems. Since extension of the methods to linear time-varying systems is straightforward, we do not discuss this matter in detail. We present only the extension of Method 2 to linear time-varying systems.

Consider the following system:

$$\overset{(n)}{x} + a_1(t)\overset{(n-1)}{x} + \cdots + a_n(t)x = b_0(t)\overset{(n)}{u} + b_1(t)\overset{(n-1)}{u} + \cdots + b_n(t)u \qquad (4\text{–}51)$$

where the $a_i(t)$ and $b_j(t)$ are differentiable as many times as we need. By defining the following set of equations:

$$\dot{x}_1 = x_2 + c_1(t)u$$
$$\dot{x}_2 = x_3 + c_2(t)u$$
$$\cdots \quad \cdots$$
$$\dot{x}_{n-1} = x_n + c_{n-1}(t)u$$
$$\dot{x}_n = -a_n(t)x_1 - a_{n-1}(t)x_2 - \cdots - a_2(t)x_{n-1} - a_1(t)x_n + c_n(t)u$$

and

$$x = x_1 + c_0(t)u \qquad (4\text{–}52)$$

we can transform Eq. (4–51) into the following form:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \cdot \\ \cdot \\ \cdot \\ \dot{x}_{n-1} \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ -a_n(t) & -a_{n-1}(t) & -a_{n-2}(t) & \cdots & -a_1(t) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_{n-1} \\ x_n \end{bmatrix} + \begin{bmatrix} c_1(t) \\ c_2(t) \\ \cdot \\ \cdot \\ \cdot \\ c_{n-1}(t) \\ c_n(t) \end{bmatrix} [u] \qquad (4\text{–}53)$$

The functions $c_1(t)$, $c_2(t)$, ..., $c_n(t)$, which are combinations of the coefficients $a_i(t)$, $b_j(t)$, and their derivatives can be determined by successively eliminating x_1, x_2, ..., x_n from Eq. (4–52) and Eq. (4–53) and requiring that the resulting nth-order scalar differential equation be identical with Eq. (4–51). For example, if $n = 2$ we obtain

$$c_0(t) = b_0(t)$$
$$c_1(t) = b_1(t) - a_1(t)b_0(t) - 2\dot{b}_0(t)$$
$$c_2(t) = b_2(t) - a_2(t)b_0(t) - a_1(t)[b_1(t) - a_1(t)b_0(t) - 2\dot{b}_0(t)] - \dot{b}_1(t) + \dot{a}_1(t)b_0(t) + \ddot{b}_0(t)$$

Similarly, the $c_i(t)$ for the general case can be obtained as follows:

$$c_0(t) = b_0(t)$$
$$c_i(t) = b_i(t) - \sum_{k=0}^{i-1}\sum_{s=0}^{i-k}\frac{(n+s-i)!}{s!(n-i)!}a_{i-k-s}(t)\overset{(s)}{c_k}(t)$$

Notice that because of the presence of derivatives of $a_i(t)$ and $b_j(t)$, the expressions for the state variables $x_1(t)$, $x_2(t)$, ..., $x_n(t)$ become complicated except in simple cases where n is small. In the case where $n = 2$ we have

$$x_1 = x - c_0(t)u$$
$$x_2 = \dot{x} - c_0(t)\dot{u} - c_1(t)u - \dot{c}_0(t)u$$

In terms of the $a_i(t)$ and $b_j(t)$, x_1 and x_2 are obtained as follows:

$$x_1 = x - b_0(t)u$$
$$x_2 = \dot{x} - b_0(t)\dot{u} - b_1(t)u + a_1(t)b_0(t)u + \dot{b}_0(t)u$$

As in the case of linear time-invariant systems, different sets of state variables can be obtained if different methods are used for deriving possible sets of state variables.

4-3 TRANSFORMING STATE SPACE EQUATIONS INTO SCALAR DIFFERENTIAL EQUATIONS

In this section we shall present techniques for deriving scalar differential equations which correspond to given state space equations. We shall show that for a given state space equation or vector matrix differential equation

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u \tag{4-54}$$

where $\mathbf{x}$ is an n vector, u is a scalar, $\mathbf{A}$ is an $n \times n$ constant matrix, and $\mathbf{B}$ is an $n \times 1$ constant matrix (an n vector), there corresponds a number of expressions of the form:

$$y^{(n)} + a_1 y^{(n-1)} + \cdots + a_{n-1}\dot{y} + a_n y = b_0 u^{(n)} + b_1 u^{(n-1)} + \cdots + b_{n-1}\dot{u} + b_n u \tag{4-55}$$

Notice, however, that in each different expression the variable y is different.

In the following we shall first demonstrate that a state space equation described by Eq. (4–54) can be transformed into an equation of the form given by Eq. (4–55). Then we shall show that by a suitable choice of the variable y Eq. (4–54) may be transformed into

$$y^{(n)} + a_1 y^{(n-1)} + \cdots + a_{n-1}\dot{y} + a_n y = u$$

Deriving scalar differential equation of the form $y^{(n)} + a_1 y^{(n-1)} + \cdots + a_{n-1}\dot{y} + a_n y = b_0 u^{(n)} + b_1 u^{(n-1)} + \cdots + b_{n-1}\dot{u} + b_n u$ from state space equation of the form $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u$ In order to derive a scalar differential equation which corresponds to a given state space equation let us first note that if the eigenvalues of $\mathbf{A}$ are distinct, or if the multiplicity of any of the multiple eigenvectors of $\mathbf{A}$ is the same as that of the corresponding multiple eigenvalue of $\mathbf{A}$, then there exists a nonsingular matrix $\mathbf{T}$ such that $\mathbf{x} = \mathbf{T}\mathbf{y}$ will transform Eq. (4–54) into

$$\dot{\mathbf{y}} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}\mathbf{y} + \mathbf{T}^{-1}\mathbf{B}u$$

where

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_2 & -a_1 \end{bmatrix}$$

and

$$\mathbf{T}^{-1}\mathbf{B} = \begin{bmatrix} c_1 \\ c_2 \\ \cdot \\ \cdot \\ \cdot \\ c_{n-1} \\ c_n \end{bmatrix}$$

The constants a_i can be determined from knowledge of the eigenvalues of $\mathbf{A}$. Once the c_i are determined by computing $\mathbf{T}^{-1}\mathbf{B}$ we can determine the b_i from Eq. (4–46).

We now give two simple examples to illustrate the procedure for deriving corresponding scalar differential equations from state space equations of the form given by Eq. (4–54).

Example 4-5 Consider the system shown in Fig. 4–3. The system can be described by

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} -3 & -2 & -1 \\ 1 & 0 & -1 \\ 0 & 0 & -4 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Define y such that it is a linear function of x_1, x_2, and x_3, or

$$y = [a \quad b \quad c] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

where a, b, and c are constants. Derive a scalar differential equation in y describing the system dynamics. (Note that y is not a unique combination of x_1, x_2, and x_3. Note also that some, but not all, of the constants a, b, and c may be zero.)

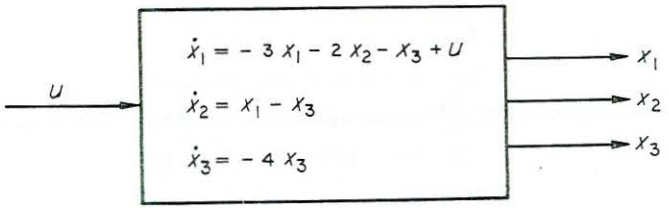


Figure 4-3 Control system.

We shall obtain the system's differential equation in the following form:

$$\dddot{y} + a_1\ddot{y} + a_2\dot{y} + a_3 y = b_0\dddot{u} + b_1\ddot{u} + b_2\dot{u} + b_3 u$$

Notice that

$$\begin{vmatrix} -3-\lambda & -2 & -1 \\ 1 & -\lambda & -1 \\ 0 & 0 & -4-\lambda \end{vmatrix} = -(\lambda^3 + 7\lambda^2 + 14\lambda + 8)$$
$$= -(\lambda+1)(\lambda+2)(\lambda+4)$$

Let us define y as the first component of the following $\mathbf{y}$:

$$\mathbf{y} = \mathbf{T}^{-1}\mathbf{x}$$

or

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

where $\mathbf{T}^{-1}$ should be chosen to be nonsingular. We determine $\mathbf{T}^{-1}$ such that

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -8 & -14 & -7 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} [u]$$

where c_1, c_2, and c_3 are constants. To determine $\mathbf{T}^{-1}$, we solve

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \begin{bmatrix} -3 & -2 & -1 \\ 1 & 0 & -1 \\ 0 & 0 & -4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -8 & -14 & -7 \end{bmatrix} \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

for a, b, c, d, e, f, g, h, and i. Note that $\mathbf{T}^{-1}$ is not uniquely determined. One possible solution (a nonsingular matrix $\mathbf{T}^{-1}$) is given by

$$\mathbf{T}^{-1} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & -5 \\ -3 & -2 & 19 \end{bmatrix}$$

For this $\mathbf{T}^{-1}$, constants c_1, c_2, and c_3 are determined as

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & -5 \\ -3 & -2 & 19 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ -3 \end{bmatrix}$$

Noting that $c_0 = 0$ because $y = y_1$, we obtain from Eq. (4–46)

$$b_0 = c_0 = 0$$
$$b_1 = c_1 + a_1 c_0 = 0$$
$$b_2 = c_2 + a_1 c_1 + a_2 c_0 = 1$$
$$b_3 = c_3 + a_1 c_2 + a_2 c_1 + a_3 c_0 = 4$$

Hence the desired scalar differential equation in y, where $y = y_1 = x_2 + x_3$, is

$$\dddot{y} + 7\ddot{y} + 14\dot{y} + 8y = \dot{u} + 4u$$

or

$$(p+1)(p+2)(p+4)y = (p+4)u, \qquad p = \frac{d}{dt}$$

Example 4-6 Given the following state space equation:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u \tag{4–56}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 3 & 0 & 2 \\ -12 & -7 & -6 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} -1 \\ 2 \\ 3 \end{bmatrix}$$

obtain the corresponding scalar differential equation of the form

$$\dddot{y} + a_1\ddot{y} + a_2\dot{y} + a_3 y = b_0\dddot{u} + b_1\ddot{u} + b_2\dot{u} + b_3 u$$

The three eigenvalues of $\mathbf{A}$ are -1, -2, and -3. We can determine a_1, a_2, and a_3 from the following equation:

$$\lambda^3 + a_1\lambda^2 + a_2\lambda + a_3 = (\lambda+1)(\lambda+2)(\lambda+3)$$

The values of a_1, a_2, and a_3 are found to be

$$a_1 = 6, \qquad a_2 = 11, \qquad a_3 = 6$$

By a suitable nonsingular transformation $\mathbf{x} = \mathbf{T}\mathbf{y}$, Eq. (4–56) can be transformed into

$$\dot{\mathbf{y}} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}\mathbf{y} + \mathbf{T}^{-1}\mathbf{B}u \tag{4–57}$$

where

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & -11 & -6 \end{bmatrix}$$

The matrix $\mathbf{T}^{-1}$ can be determined by the method discussed in Example 4–5. Here we shall, however, find $\mathbf{T}^{-1}$ by a different method. Let $\mathbf{J}$ be the Jordan canonical form of $\mathbf{A}$. Then there exists a nonsingular matrix $\mathbf{P}_1$ such that $\mathbf{P}_1^{-1}\mathbf{A}\mathbf{P}_1 = \mathbf{J}$. Note also that there exists a nonsingular matrix $\mathbf{P}_2$ such that $\mathbf{P}_2^{-1}(\mathbf{T}^{-1}\mathbf{A}\mathbf{T})\mathbf{P}_2 = \mathbf{J}$. Hence, $\mathbf{P}_1^{-1}\mathbf{A}\mathbf{P}_1 = \mathbf{P}_2^{-1}\mathbf{T}^{-1}\mathbf{A}\mathbf{T}\mathbf{P}_2$, or

$$\mathbf{A} = \mathbf{P}_1\mathbf{P}_2^{-1}\mathbf{T}^{-1}\mathbf{A}\mathbf{T}\mathbf{P}_2\mathbf{P}_1^{-1} = (\mathbf{T}\mathbf{P}_2\mathbf{P}_1^{-1})^{-1}\mathbf{A}(\mathbf{T}\mathbf{P}_2\mathbf{P}_1^{-1})$$

Therefore, we can obtain $\mathbf{T}^{-1}$ from

$$\mathbf{T}^{-1} = \mathbf{P}_2\mathbf{P}_1^{-1}$$

where

$\mathbf{P}_1$ = matrix defined by Eq. (3–18) or matrix (3–21)
$\mathbf{P}_2$ = matrix defined by Eq. (3–37)

(For multiple eigenvalue systems we must make necessary modifications to $\mathbf{P}_1$ and $\mathbf{P}_2$. Notice that a nonsingular matrix $\mathbf{T}$ does not exist unless the multiplicity of any of the multiple eigenvectors of $\mathbf{A}$ is the same as that of the corresponding multiple eigenvalue of $\mathbf{A}$.)

For the present system $\mathbf{P}_2$ and $\mathbf{P}_1^{-1}$ are (Refer to Prob. A-3-1.):

$$\mathbf{P}_2 = \begin{bmatrix} 1 & 1 & 1 \\ -1 & -2 & -3 \\ 1 & 4 & 9 \end{bmatrix}, \qquad \mathbf{P}_1^{-1} = \begin{bmatrix} \frac{9}{2} & \frac{5}{2} & 1 \\ -3 & -2 & -1 \\ \frac{5}{2} & \frac{3}{2} & 1 \end{bmatrix}$$

Equation (4–57) can now be written as

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & -11 & -6 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + \begin{bmatrix} 3 \\ -6 \\ 19 \end{bmatrix} [u]$$

Note that $y = y_1 + c_0 u$. By use of Eq. (4–46) we obtain

$$\begin{aligned} b_0 &= c_0 \\ b_1 &= c_1 + a_1 c_0 = 3 + 6c_0 \\ b_2 &= c_2 + a_1 c_1 + a_2 c_0 = 12 + 11c_0 \\ b_3 &= c_3 + a_1 c_2 + a_2 c_1 + a_3 c_0 = 16 + 6c_0 \end{aligned}$$

Therefore, a scalar differential equation equivalent to Eq. (4–56) is given by

$$\dddot{y} + 6\ddot{y} + 11\dot{y} + 6y = c_0\dddot{u} + (3 + 6c_0)\ddot{u} + (12 + 11c_0)\dot{u} + (16 + 6c_0)u$$

or

$$\frac{d^3}{dt^3}(y - c_0 u) + 6\frac{d^2}{dt^2}(y - c_0 u) + 11\frac{d}{dt}(y - c_0 u) + 6(y - c_0 u) = 3\ddot{u} + 12\dot{u} + 16u$$

Here, c_0 is an arbitrary constant. In terms of y_1, this scalar differential equation becomes

$$\dddot{y}_1 + 6\ddot{y}_1 + 11\dot{y}_1 + 6y_1 = 3\ddot{u} + 12\dot{u} + 16u$$

Notice that y_1 and $\mathbf{x}$ are related by the following equation:

$$y_1 = 4x_1 + 2x_2 + x_3$$

If we choose a different quantity as the dependent variable in the scalar differential equation, then the resulting scalar differential equation is, of course, different. In fact, if we make the following transformation:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -1 & 2 & 3 \\ 2 & 3 & -34 \\ 3 & -20 & 75 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

then the scalar differential equation (in terms of z_3) corresponding to Eq. (4–56) becomes

$$\dddot{z}_3 + 6\ddot{z}_3 + 11\dot{z}_3 + 6z_3 = u$$

(Refer to Prob. A-4-10.) This is the simplest form of the scalar differential equation we can obtain for the present state space equation. Such a transformation into the simplest form may be made if we make a suitable choice of the dependent variable of the scalar differential equation.

In the following we shall discuss a method for transforming a state space equation into the simplest scalar differential equation.

Deriving scalar differential equation of the form $y^{(n)} + a_1 y^{(n-1)} + \cdots + a_{n-1}\dot{y} + a_n y = u$ from state space equation of the form $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u$ (4–1) Consider the system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u \tag{4–58}$$

where $\mathbf{x}$ is an n vector, u is a scalar, $\mathbf{A}$ is an $n \times n$ constant matrix, and $\mathbf{B}$ is an $n \times 1$ constant matrix. We shall show that by means of a suitable transformation

$$\mathbf{x} = \mathbf{H}\mathbf{y} \qquad (\mathbf{H}\text{: nonsingular})$$

we may be able to transform Eq. (4–58) into the following form:

$$\dot{\mathbf{y}} = \mathbf{H}^{-1}\mathbf{A}\mathbf{H}\mathbf{y} + \mathbf{H}^{-1}\mathbf{B}u \tag{4–59}$$

where

$$\mathbf{H}^{-1}\mathbf{A}\mathbf{H} = \begin{bmatrix} 0 & 0 & \cdots & 0 & -a_n \\ 1 & 0 & \cdots & 0 & -a_{n-1} \\ 0 & 1 & \cdots & 0 & -a_{n-2} \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & \cdots & 1 & -a_1 \end{bmatrix}, \qquad \mathbf{H}^{-1}\mathbf{B} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

and

$$\begin{aligned} -a_n &= (\mathbf{h}_1, \mathbf{A}^n\mathbf{B}) \\ -a_{n-1} &= (\mathbf{h}_2, \mathbf{A}^n\mathbf{B}) \\ &\cdots \qquad \cdots \\ -a_1 &= (\mathbf{h}_n, \mathbf{A}^n\mathbf{B}) \end{aligned} \tag{4–60}$$

In Eq. (4–60) $\mathbf{h}_i$ is the ith row of $\mathbf{H}^{-1}$ and $(\mathbf{h}_i, \mathbf{A}^n\mathbf{B})$ denotes the inner product of $\mathbf{h}_i$ and $\mathbf{A}^n\mathbf{B}$. By eliminating $y_1, y_2, \ldots, y_{n-1}$ from Eq. (4–59) we obtain

$$y_n^{(n)} + a_1 y_n^{(n-1)} + a_2 y_n^{(n-2)} + \cdots + a_{n-1}\dot{y}_n + a_n y_n = u \tag{4–61}$$

It can be shown that such a transformation matrix $\mathbf{H}$ is given by a matrix consisting of column vectors $\mathbf{B}, \mathbf{AB}, \mathbf{A}^2\mathbf{B}, \ldots, \mathbf{A}^{n-1}\mathbf{B}$, or

$$\mathbf{H} = [\mathbf{B} \mid \mathbf{AB} \mid \mathbf{A}^2\mathbf{B} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{B}]$$

provided n column vectors are linearly independent. (See Prob. A-4-9.)

Notice that if we define

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ \cdot \\ \cdot \\ \cdot \\ z_n \end{bmatrix} = \begin{bmatrix} y_n \\ \dot{y}_n \\ \cdot \\ \cdot \\ \cdot \\ y_n^{(n-1)} \end{bmatrix}$$

then Eq. (4–59) can be transformed into

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \cdot \\ \cdot \\ \cdot \\ \dot{z}_{n-1} \\ \dot{z}_n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \cdot \\ \cdot \\ \cdot \\ z_{n-1} \\ z_n \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix} [u] \tag{4–62}$$

This means that Eq. (4–58) may be transformed into Eq. (4–62), or its equivalence

$$\overset{(n)}{z_1} + a_1 \overset{(n-1)}{z_1} + \cdots + a_{n-1}\dot{z}_1 + a_n z_1 = u$$

Example 4-7 Consider the following system:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} -6 & 1 & 0 \\ -11 & 0 & 1 \\ -6 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 2 \\ 6 \\ 2 \end{bmatrix} [u]$$

We shall illustrate that this state space equation can be transformed into the scalar differential equation

$$\dddot{y} + a_1\ddot{y} + a_2\dot{y} + a_3 y = u$$

Since for this system

$$\mathbf{B} = \begin{bmatrix} 2 \\ 6 \\ 2 \end{bmatrix}, \qquad \mathbf{AB} = \begin{bmatrix} -6 \\ -20 \\ -12 \end{bmatrix}, \qquad \mathbf{A^2B} = \begin{bmatrix} 16 \\ 54 \\ 36 \end{bmatrix}$$

we obtain

$$\mathbf{H} = \begin{bmatrix} 2 & -6 & 16 \\ 6 & -20 & 54 \\ 2 & -12 & 36 \end{bmatrix}$$

Let

$$\mathbf{x} = \mathbf{H}\mathbf{y}$$

then we obtain

$$\dot{\mathbf{y}} = \mathbf{H}^{-1}\mathbf{A}\mathbf{H}\mathbf{y} + \mathbf{H}^{-1}\mathbf{B}u$$

where

$$\mathbf{H}^{-1}\mathbf{A}\mathbf{H} = \begin{bmatrix} 0 & 0 & -6 \\ 1 & 0 & -11 \\ 0 & 1 & -6 \end{bmatrix}$$

and

$$\mathbf{H}^{-1}\mathbf{B} = \begin{bmatrix} 9 & -3 & \frac{1}{2} \\ \frac{27}{2} & -5 & \frac{3}{2} \\ 4 & -\frac{3}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 2 \\ 6 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Rewriting

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & -6 \\ 1 & 0 & -11 \\ 0 & 1 & -6 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} [u]$$

Elimination of y_1 and y_2 from the foregoing equation results in

$$\dddot{y}_3 + 6\ddot{y}_3 + 11\dot{y}_3 + 6y_3 = u$$

where y_3 is related to x_1, x_2, and x_3 by the following equation:

$$y_3 = 4x_1 - \frac{3}{2}x_2 + \frac{1}{2}x_3$$

4-4 PARTIAL-FRACTION EXPANSION TECHNIQUE

(*4–2, 4–3*)

The partial-fraction expansion technique can be used for obtaining state space representation of transfer functions or scalar differential equations. By this technique we can obtain directly the state space equation in which the coefficient matrix of the state vector is in the Jordan canonical form. In applying the partial-fraction expansion technique to the transfer function, it is necessary to factor the denominator.

The procedure for obtaining a state space equation for the differential equation

$$\overset{(n)}{x} + a_1\overset{(n-1)}{x} + \cdots + a_{n-1}\dot{x} + a_n x = b_0\overset{(n)}{u} + b_1\overset{(n-1)}{u} + \cdots + b_{n-1}\dot{u} + b_n u \tag{4–63}$$

by the partial-fraction expansion technique is to assume that the initial conditions are zero and write the corresponding transfer function. The assumption of zero initial conditions is permitted because we are dealing with linear systems, and state space representation of linear differential systems does not depend on the initial conditions. A state space equation for the differential equation with the zero initial conditions (or the corresponding transfer function) is also a state space equation for the differential equation with the nonzero initial conditions.

In this book, therefore, whenever we use the partial-fraction expansion technique we first assume the zero initial conditions in the differential

equation and obtain the transfer function. Then we derive the corresponding state space equation. Finally, we relax the zero initial conditions to arbitrary initial conditions.

System with simple forcing function We shall first consider a simple case where Eq. (4–63) assumes the following form:

$$\overset{(n)}{x} + a_1 \overset{(n-1)}{x} + \cdots + a_{n-1}\dot{x} + a_n x = bu$$

The transfer function corresponding to this equation is

$$\frac{X(s)}{U(s)} = \frac{b}{(s-\lambda_1)(s-\lambda_2)\cdots(s-\lambda_n)}$$

We shall transform this transfer function into a first-order vector matrix differential equation, or a state space equation. We shall first consider the transfer function with distinct poles and then consider that with multiple poles.

State space representation of transfer function with distinct poles Let us denote the transfer function as $W(s)$ or

$$W(s) = \frac{X(s)}{U(s)} = \frac{b}{s^n + a_1 s^{n-1} + \cdots + a_{n-1}s + a_n} = \frac{b}{(s-\lambda_1)(s-\lambda_2)\cdots(s-\lambda_n)} \tag{4–64}$$

We assume here that the poles $\lambda_1, \lambda_2, \ldots, \lambda_n$ are distinct. The original scalar differential equation from which Eq. (4–64) is obtained can be represented in a state space by Eq. (4–8) or Eq. (4–9). If each of Eq. (4–9) is Laplace transformed assuming zero initial conditions, then we obtain using variables $y_1, y_2, \ldots, y_n$ rather than variables $z_1, z_2, \ldots, z_n$

$$\begin{aligned} \frac{Y_1(s)}{U(s)} &= \frac{1}{s-\lambda_1} \\ \frac{Y_2(s)}{U(s)} &= \frac{1}{s-\lambda_2} \\ &\cdots \\ \frac{Y_n(s)}{U(s)} &= \frac{1}{s-\lambda_n} \end{aligned} \tag{4–65}$$

Hence we see that a simplified[1] state space representation of Eq. (4–64) is

[1]Here, *simplified state space representation* means that, in the state space equation, the coefficient matrix of the state vector is in the Jordan canonical form (which includes the diagonal form as a special case) and the elements of the coefficient matrix of the forcing function u are either unity or zero.

essentially a representation of the transfer function by n simultaneous first-order transfer functions which can be obtained by dividing Eq. (4–64) into n first-order systems, as shown in Fig. 4–4. Since Eq. (4–64) and its state space representation, Eq. (4–65), are equivalent, the former can be derived from the latter, and vice versa. Essentially, Eq. (4–64) and Eq. (4–65) are related by a transformation matrix $\mathbf{P}$. We can find the necessary transformation matrix $\mathbf{P}$ by expanding the transfer function into partial-fractions.

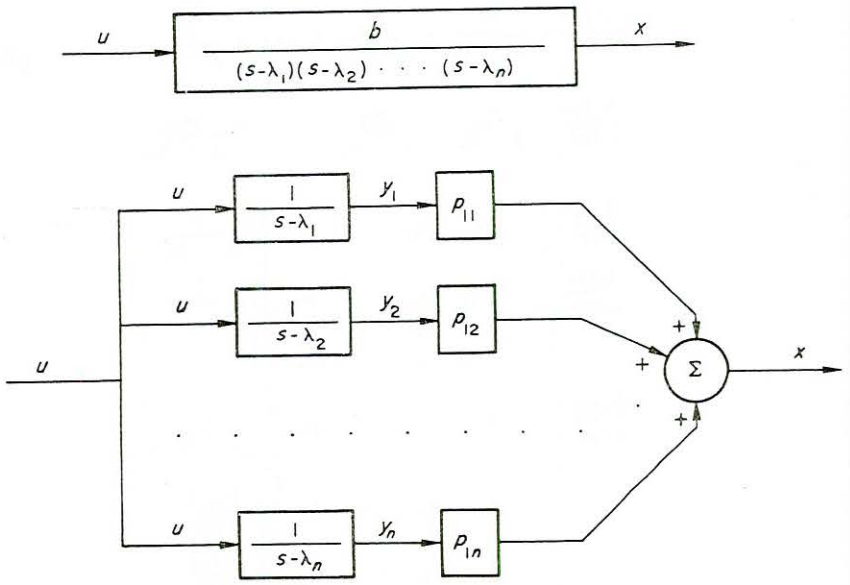


Figure 4-4 An nth-order transfer function and its representation by a sum of n first-order transfer functions.

Let us define

$$\begin{aligned} X_1(s) &= X(s) \\ X_2(s) &= sX(s) \\ &\cdots \\ X_n(s) &= s^{n-1}X(s) \end{aligned}$$

If the following transformation is made

$$\mathbf{X}(s) = \mathbf{P}\mathbf{Y}(s)$$

then

$$\frac{\mathbf{X}(s)}{U(s)} = \mathbf{P}\frac{\mathbf{Y}(s)}{U(s)} \tag{4–66}$$

Since

$$\frac{X_1(s)}{U(s)} = W(s)$$

$$\frac{X_2(s)}{U(s)} = sW(s)$$

$$\cdots$$

$$\frac{X_n(s)}{U(s)} = s^{n-1}W(s)$$

if each equation is expanded into partial-fractions, then

$$\begin{aligned}
\frac{X_1(s)}{U(s)} &= \frac{p_{11}}{s-\lambda_1} + \frac{p_{12}}{s-\lambda_2} + \cdots + \frac{p_{1n}}{s-\lambda_n} \\
\frac{X_2(s)}{U(s)} &= \frac{p_{21}}{s-\lambda_1} + \frac{p_{22}}{s-\lambda_2} + \cdots + \frac{p_{2n}}{s-\lambda_n} \\
&\cdots \qquad\qquad \cdots \qquad\qquad \cdots \\
\frac{X_n(s)}{U(s)} &= \frac{p_{n1}}{s-\lambda_1} + \frac{p_{n2}}{s-\lambda_2} + \cdots + \frac{p_{nn}}{s-\lambda_n}
\end{aligned} \tag{4-67}$$

or

$$\begin{bmatrix} \frac{X_1(s)}{U(s)} \\ \frac{X_2(s)}{U(s)} \\ \vdots \\ \frac{X_n(s)}{U(s)} \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} \begin{bmatrix} \frac{1}{s-\lambda_1} \\ \frac{1}{s-\lambda_2} \\ \vdots \\ \frac{1}{s-\lambda_n} \end{bmatrix}$$

$$= \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} \begin{bmatrix} \frac{Y_1(s)}{U(s)} \\ \frac{Y_2(s)}{U(s)} \\ \vdots \\ \frac{Y_n(s)}{U(s)} \end{bmatrix} \tag{4-68}$$

From Eq. (4–68) the transformation matrix $\mathbf{P}$ is obtained as

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} \tag{4-69}$$

It is seen that each element of $\mathbf{P}$ is obtained directly from the coefficients of the partial-fraction expansions of $W(s)$, $sW(s)$, . . . , $s^{n-1}W(s)$. The simplified state space representation of Eq. (4–64) then becomes

$$\dot{\mathbf{y}} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}\mathbf{y} + \mathbf{F}u$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \\ -a_n & -a_{n-1} & \cdots & -a_1 \end{bmatrix}, \qquad \mathbf{F} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}$$

or

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \vdots \\ \dot{y}_n \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} [u]$$

It is noted that the preceding form is the same regardless of the value of b in Eq. (4–64). (The elements of $\mathbf{P}$, however, depend on the value of b.)

If the first column of the matrix $\mathbf{P}$ given by Eq. (4–69) is divided by p_{11}, the second column by p_{12}, . . . , the last column by p_{1n}, then the transformation matrix $\mathbf{P}$ becomes $\mathbf{P}_1$, given by

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \frac{p_{21}}{p_{11}} & \frac{p_{22}}{p_{12}} & \cdots & \frac{p_{2n}}{p_{1n}} \\ \frac{p_{31}}{p_{11}} & \frac{p_{32}}{p_{12}} & \cdots & \frac{p_{3n}}{p_{1n}} \\ \vdots & \vdots & & \vdots \\ \frac{p_{n1}}{p_{11}} & \frac{p_{n2}}{p_{12}} & \cdots & \frac{p_{nn}}{p_{1n}} \end{bmatrix}$$

This matrix must be equal to

$$\mathbf{P} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_n \\ \lambda_1^2 & \lambda_2^2 & \cdots & \lambda_n^2 \\ \vdots & \vdots & & \vdots \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \cdots & \lambda_n^{n-1} \end{bmatrix}$$

Hence

$$\frac{p_{21}}{p_{11}} = \lambda_1, \qquad \frac{p_{22}}{p_{12}} = \lambda_2, \ \ldots, \ \frac{p_{2n}}{p_{1n}} = \lambda_n$$

or in a general form

$$p_{ij} = p_{1j}\lambda_j^{i-1} \tag{4–70}$$

Thus, once $p_{11}, p_{12}, \ldots, p_{1n}$ are evaluated, $p_{21}, p_{22}, \ldots, p_{nn}$ can be obtained from Eq. (4–70). [Notice that Eq. (4–70) applies only to the case of distinct eigenvalues.]

Example 4-8 Consider the following transfer function:

$$\frac{X(s)}{U(s)} = W(s) = \frac{1}{(s+1)(s+2)(s+3)}$$

By use of the partial-fraction expansion technique obtain a state space representation of this transfer function.

Since

$$\frac{X_1(s)}{U(s)} = W(s) = \frac{1}{(s+1)(s+2)(s+3)}$$

$$\frac{X_2(s)}{U(s)} = sW(s) = \frac{s}{(s+1)(s+2)(s+3)}$$

$$\frac{X_3(s)}{U(s)} = s^2W(s) = \frac{s^2}{(s+1)(s+2)(s+3)}$$

the partial-fraction expansions of these three equations become

$$\frac{X_1(s)}{U(s)} = \frac{1/2}{s+1} + \frac{-1}{s+2} + \frac{1/2}{s+3}$$

$$\frac{X_2(s)}{U(s)} = \frac{-1/2}{s+1} + \frac{2}{s+2} + \frac{-3/2}{s+3}$$

$$\frac{X_3(s)}{U(s)} = \frac{1/2}{s+1} + \frac{-4}{s+2} + \frac{9/2}{s+3}$$

or

$$\begin{bmatrix} \frac{X_1(s)}{U(s)} \\ \frac{X_2(s)}{U(s)} \\ \frac{X_3(s)}{U(s)} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -1 & \frac{1}{2} \\ -\frac{1}{2} & 2 & -\frac{3}{2} \\ \frac{1}{2} & -4 & \frac{9}{2} \end{bmatrix} \begin{bmatrix} \frac{1}{s+1} \\ \frac{1}{s+2} \\ \frac{1}{s+3} \end{bmatrix}$$

Thus

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2} & -1 & \frac{1}{2} \\ -\frac{1}{2} & 2 & -\frac{3}{2} \\ \frac{1}{2} & -4 & \frac{9}{2} \end{bmatrix}$$

If we make the transformation $\mathbf{x} = \mathbf{P}\mathbf{y}$, or

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -1 & \frac{1}{2} \\ -\frac{1}{2} & 2 & -\frac{3}{2} \\ \frac{1}{2} & -4 & \frac{9}{2} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \tag{4–71}$$

then

$$\begin{bmatrix} \frac{Y_1(s)}{U(s)} \\ \frac{Y_2(s)}{U(s)} \\ \frac{Y_3(s)}{U(s)} \end{bmatrix} = \begin{bmatrix} \frac{1}{s+1} \\ \frac{1}{s+2} \\ \frac{1}{s+3} \end{bmatrix}$$

Hence the given transfer function can be transformed into a set of three simultaneous equations as follows:

$$\begin{aligned} \dot{y}_1 &= -y_1 + u \\ \dot{y}_2 &= -2y_2 + u \\ \dot{y}_3 &= -3y_3 + u \end{aligned} \tag{4–72}$$

This is the simplified state space representation of the given transfer function. Here the y_i and x_i $(i = 1, 2, 3)$ are related by Eq. (4–71) and $x_1 = x$, $x_2 = \dot{x}$, $x_3 = \ddot{x}$. The response $x(t)$ is clearly given by

$$x(t) = \frac{1}{2}y_1(t) - y_2(t) + \frac{1}{2}y_3(t)$$

where $y_1(t)$, $y_2(t)$, and $y_3(t)$ can be easily obtained from Eq. (4–72) once the forcing function u is specified. The initial condition terms are given by

$$\mathbf{y}(0) = \mathbf{P}^{-1}\mathbf{x}(0)$$

Notice that Eq. (4–72) is also a state space representation of the following differential equation:

$$(p+1)(p+2)(p+3)x = u, \qquad p = \frac{d}{dt}$$

or

$$\dddot{x} + 6\ddot{x} + 11\dot{x} + 6x = u$$

State space representation of transfer function with multiple poles Next, consider the case where some poles of the transfer function are multiple poles. Consider the following transfer function:

$$\begin{aligned} \frac{X(s)}{U(s)} = W(s) &= \frac{b}{s^n + a_1s^{n-1} + \cdots + a_{n-1}s + a_n} \\ &= \frac{b}{(s-\lambda_1)^2(s-\lambda_3)\cdots(s-\lambda_n)} \end{aligned} \tag{4–73}$$

This transfer function has double poles. The simplified state space equations

corresponding to the double poles should be in the form

$$\dot{y}_1 = \lambda_1 y_1 + y_2$$
$$\dot{y}_2 = \lambda_1 y_2 + u$$

from which

$$\frac{Y_1(s)}{Y_2(s)} = \frac{1}{s - \lambda_1}$$

This suggests that $\mathbf{Y}(s)/U(s)$ be defined by

$$\begin{aligned}
\frac{Y_1(s)}{U(s)} &= \frac{1}{(s-\lambda_1)^2} \\
\frac{Y_2(s)}{U(s)} &= \frac{1}{s-\lambda_1} \\
\frac{Y_3(s)}{U(s)} &= \frac{1}{s-\lambda_3} \\
&\cdots \quad \cdots \\
\frac{Y_n(s)}{U(s)} &= \frac{1}{s-\lambda_n}
\end{aligned} \tag{4–74}$$

Let $X_1(s) = X(s)$, $X_2(s) = sX(s)$, $\ldots$, $X_n(s) = s^{n-1}X(s)$. If

$$\frac{X_1(s)}{U(s)} = W(s), \qquad \frac{X_2(s)}{U(s)} = sW(s), \quad \cdots, \quad \frac{X_n(s)}{U(s)} = s^{n-1}W(s)$$

are expanded into partial-fractions, then

$$\begin{aligned}
\frac{X_1(s)}{U(s)} &= \frac{p_{11}}{(s-\lambda_1)^2} + \frac{p_{12}}{s-\lambda_1} + \frac{p_{13}}{s-\lambda_3} + \cdots + \frac{p_{1n}}{s-\lambda_n} \\
\frac{X_2(s)}{U(s)} &= \frac{p_{21}}{(s-\lambda_1)^2} + \frac{p_{22}}{s-\lambda_1} + \frac{p_{23}}{s-\lambda_3} + \cdots + \frac{p_{2n}}{s-\lambda_n} \\
&\cdots \quad \cdots \quad \cdots \\
\frac{X_n(s)}{U(s)} &= \frac{p_{n1}}{(s-\lambda_1)^2} + \frac{p_{n2}}{s-\lambda_1} + \frac{p_{n3}}{s-\lambda_3} + \cdots + \frac{p_{nn}}{s-\lambda_n}
\end{aligned} \tag{4–75}$$

Let

$$\mathbf{P}_p = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} \tag{4–76}$$

By dividing the second column of $\mathbf{P}_p$ by a constant factor

$$\frac{b}{(\lambda_1 - \lambda_3)(\lambda_1 - \lambda_4)\cdots(\lambda_1 - \lambda_n)}$$

and by dividing the first column, third column, $\ldots$, nth column by p_{11}, $p_{13}, \ldots, p_{1n}$, respectively, $\mathbf{P}_p$ can be modified into $\mathbf{P}_1$ given by:

$$\mathbf{P}_1 = \begin{bmatrix} 1 & -L & 1 & \cdots & 1 \\ \lambda_1 & 1-\lambda_1 L & \lambda_3 & \cdots & \lambda_n \\ \lambda_1^2 & 2\lambda_1 - \lambda_1^2 L & \lambda_3^2 & \cdots & \lambda_n^2 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \lambda_1^{n-1} & (n-1)\lambda_1^{n-2} - \lambda_1^{n-1}L & \lambda_3^{n-1} & \cdots & \lambda_n^{n-1} \end{bmatrix} \tag{4–77}$$

where

$$L = \frac{1}{\lambda_1 - \lambda_3} + \frac{1}{\lambda_1 - \lambda_4} + \cdots + \frac{1}{\lambda_1 - \lambda_n}$$

Then

$$\mathbf{P}_1 = \mathbf{P} + \begin{bmatrix} 0 & -L & \cdots & 0 \\ 0 & -\lambda_1 L & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & -\lambda_1^{n-1}L & \cdots & 0 \end{bmatrix}$$

where $\mathbf{P}$ is defined by Eq. (3–50). Also, $\mathbf{P}_1^{-1}$ is obtained as

$$\mathbf{P}_1^{-1} = \mathbf{P}^{-1} + \begin{bmatrix} \frac{L\Delta_1}{|\mathbf{P}|} & \frac{L\Delta_2}{|\mathbf{P}|} & \cdots & \frac{L\Delta_n}{|\mathbf{P}|} \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

(Refer to Prob. A-4-4.) Thus,

$$\begin{aligned}
\mathbf{P}_1^{-1}\mathbf{A}\mathbf{P}_1 &= \mathbf{P}^{-1}\mathbf{A}\mathbf{P} \\
&+ \begin{bmatrix} \frac{L\Delta_1}{|\mathbf{P}|} & \frac{L\Delta_2}{|\mathbf{P}|} & \cdots & \frac{L\Delta_n}{|\mathbf{P}|} \\ 0 & 0 & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & 0 \end{bmatrix} \mathbf{A}\mathbf{P} + \mathbf{P}^{-1}\mathbf{A} \begin{bmatrix} 0 & -L & \cdots & 0 \\ 0 & -\lambda_1 L & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & -\lambda_1^{n-1}L & \cdots & 0 \end{bmatrix} \\
&+ \begin{bmatrix} \frac{L\Delta_1}{|\mathbf{P}|} & \frac{L\Delta_2}{|\mathbf{P}|} & \cdots & \frac{L\Delta_n}{|\mathbf{P}|} \\ 0 & 0 & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & 0 \end{bmatrix} \mathbf{A} \begin{bmatrix} 0 & -L & \cdots & 0 \\ 0 & -\lambda_1 L & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & -\lambda_1^{n-1}L & \cdots & 0 \end{bmatrix}
\end{aligned} \tag{4–78}$$

where

$$\mathrm{A} = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & 1 \\ -a_n & -a_{n-1} & \cdots & -a_1 \end{bmatrix}$$

The second and third terms of the right-hand side of Eq. (4–78) will cancel each other and the last term of the right-hand side of Eq. (4–78) is zero. (Refer to Prob. A-4-5.) Thus, we find $\mathbf{P}_1^{-1}\mathbf{A}\mathbf{P}_1 = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$. This implies that

$$\mathbf{P}_p^{-1}\mathbf{A}\mathbf{P}_p = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$$

This means that as a transformation matrix $\mathbf{P}$ we may use either Eq. (3–50) or Eq. (4–76) or Eq. (4–77).

The simplified state space equation of the system with double eigenvalues is given by

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \\ \cdot \\ \cdot \\ \cdot \\ \dot{y}_n \end{bmatrix} = \begin{bmatrix} \lambda_1 & 1 & 0 & \cdots & 0 \\ 0 & \lambda_1 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix} [u] \qquad (4\text{–}79)$$

This result is the same for any value of b in Eq. (4–73).

Similar results can be obtained for the case where the transfer function involves multiple poles of multiplicity m. In this case $\mathbf{Y}(s)/U(s)$ are defined by

$$\begin{aligned} \frac{Y_1(s)}{U(s)} &= \frac{1}{(s-\lambda_1)^m} \\ \frac{Y_2(s)}{U(s)} &= \frac{1}{(s-\lambda_1)^{m-1}} \\ &\cdots \quad \cdots \\ \frac{Y_m(s)}{U(s)} &= \frac{1}{s-\lambda_1} \\ \frac{Y_{m+1}(s)}{U(s)} &= \frac{1}{s-\lambda_{m+1}} \\ &\cdots \quad \cdots \\ \frac{Y_n(s)}{U(s)} &= \frac{1}{s-\lambda_n} \end{aligned} \qquad (4\text{–}80)$$

Notice that

$$\frac{Y_1(s)}{Y_2(s)} = \frac{1}{s-\lambda_1}, \qquad \frac{Y_2(s)}{Y_3(s)} = \frac{1}{s-\lambda_1}, \quad \cdots, \quad \frac{Y_{m-1}(s)}{Y_m(s)} = \frac{1}{s-\lambda_1}$$

The simplified state space equation is

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \\ \cdot \\ \cdot \\ \cdot \\ \dot{y}_m \\ \dot{y}_{m+1} \\ \cdot \\ \cdot \\ \cdot \\ \dot{y}_n \end{bmatrix} = \left[\begin{array}{ccccc:cc} \lambda_1 & 1 & 0 & \cdots & 0 & & 0 \\ 0 & \lambda_1 & 1 & \cdots & 0 & & \\ 0 & 0 & \lambda_1 & \cdots & 0 & & \\ \cdot & \cdot & \cdot & & \cdot & & \\ \cdot & \cdot & \cdot & & \cdot & & \\ \cdot & \cdot & \cdot & & \cdot & & \\ 0 & 0 & 0 & \cdots & \lambda_1 & & \\ \hdashline & & & & & \lambda_{m+1} & 0 \\ & & & & & \cdot & \\ & & & & & & \cdot \\ 0 & & & & & 0 & \lambda_n \end{array}\right] \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \cdot \\ \cdot \\ \cdot \\ y_m \\ y_{m+1} \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix} [u] \qquad (4\text{–}81)$$

The transformation matrix $\mathbf{P}$ can be obtained from the coefficients of partial-fraction expansions of

$$\frac{X_1(s)}{U(s)}, \qquad \frac{X_2(s)}{U(s)}, \quad \cdots, \quad \frac{X_n(s)}{U(s)}$$

It is similar to that given by Eq. (4–76), where the first m elements $p_{i1}, p_{i2}, \ldots, p_{im}$ $(i = 1, 2, 3, \ldots, n)$ in each row correspond to the m multiple poles.

Example 4-9 Consider the following transfer function:

$$\frac{X(s)}{U(s)} = W(s) = \frac{5}{(s+1)^2(s+2)}$$

Obtain a state space equation for this transfer function.

Since

$$\begin{aligned} \frac{X_1(s)}{U(s)} &= \frac{5}{(s+1)^2} + \frac{-5}{s+1} + \frac{5}{s+2} \\ \frac{X_2(s)}{U(s)} &= \frac{-5}{(s+1)^2} + \frac{10}{s+1} + \frac{-10}{s+2} \\ \frac{X_3(s)}{U(s)} &= \frac{5}{(s+1)^2} + \frac{-15}{s+1} + \frac{20}{s+2} \end{aligned}$$

the transformation matrix $\mathbf{P}_p$ defined by Eq. (4–76) is obtained as

$$\mathbf{P}_p = \begin{bmatrix} 5 & -5 & 5 \\ -5 & 10 & -10 \\ 5 & -15 & 20 \end{bmatrix} \qquad (4\text{–}82)$$

By means of the transformation $\mathbf{x} = \mathbf{P}_p\mathbf{y}$ the given transfer function can be transformed into the following simplified state space equation:

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} [u] \tag{4-83}$$

From this analysis we can immediately obtain a state space equation for the differential equation

$$(p+1)^2(p+2)x = 5u, \qquad p = \frac{d}{dt}$$

or

$$\dddot{x} + 4\ddot{x} + 5\dot{x} + 2x = 5u$$

The simplified form of the state space equation for this differential equation is given by Eq. (4–83). The initial condition terms are given by

$$\begin{bmatrix} y_1(0) \\ y_2(0) \\ y_3(0) \end{bmatrix} = \begin{bmatrix} 5 & -5 & 5 \\ -5 & 10 & -10 \\ 5 & -15 & 20 \end{bmatrix}^{-1} \begin{bmatrix} x(0) \\ \dot{x}(0) \\ \ddot{x}(0) \end{bmatrix}$$

and x is given by

$$x = 5y_1 - 5y_2 + 5y_3$$

System with forcing terms involving derivatives of forcing function u We shall next treat the system equation of the form

$$\overset{(n)}{x} + a_1\overset{(n-1)}{x} + \cdots + a_{n-1}\dot{x} + a_n x = b_0\overset{(n)}{u} + b_1\overset{(n-1)}{u} + \cdots + b_{n-1}\dot{u} + b_n u$$

In order to simplify our discussion, we first consider the case where $n = 3$. Then, the differential equation becomes

$$\dddot{x} + a_1\ddot{x} + a_2\dot{x} + a_3 x = b_0\dddot{u} + b_1\ddot{u} + b_2\dot{u} + b_3 u$$

The corresponding transfer function is

$$\frac{X(s)}{U(s)} = \frac{b_0s^3 + b_1s^2 + b_2s + b_3}{s^3 + a_1s^2 + a_2s + a_3} = \frac{b_0s^3 + b_1s^2 + b_2s + b_3}{(s-\lambda_1)(s-\lambda_2)(s-\lambda_3)}$$

Here we assume that the poles λ_1, λ_2, and λ_3 are distinct. (Extension to the case of multiple poles is straightforward.) We cannot use Eq. (4–67), since the order of the numerator of any of $X_i(s)/U(s)$ $(i = 1, 2, \ldots, n)$ is equal to or will exceed that of the corresponding denominator and, therefore, the partial-fraction expansion approach must be modified.

Consider, as an example, the following differential equation:

$$\dddot{x} + 6\ddot{x} + 11\dot{x} + 6x = \dddot{u} + 8\ddot{u} + 17\dot{u} + 8u \tag{4-84}$$

The corresponding transfer function is

$$\frac{X(s)}{U(s)} = \frac{s^3 + 8s^2 + 17s + 8}{(s+1)(s+2)(s+3)} \tag{4-85}$$

From the analysis in Section 4–2, we know that the simplified state space representation of the system can be in the following form:

$$\frac{Y_1(s)}{U(s)} = \frac{1}{s+1}$$
$$\frac{Y_2(s)}{U(s)} = \frac{1}{s+2} \tag{4-86}$$
$$\frac{Y_3(s)}{U(s)} = \frac{1}{s+3}$$

In order to find $Y_1(s)$, $Y_2(s)$, and $Y_3(s)$, let us divide numerators of the right-hand sides of $X(s)/U(s)$, $sX(s)/U(s)$, and $s^2X(s)/U(s)$ by respective denominators and expand remaining terms into partial fractions.

$$\frac{X(s)}{U(s)} = 1 - \frac{1}{s+1} + \frac{2}{s+2} + \frac{1}{s+3}$$
$$\frac{sX(s)}{U(s)} = s + 2 + \frac{1}{s+1} - \frac{4}{s+2} - \frac{3}{s+3}$$
$$\frac{s^2X(s)}{U(s)} = s^2 + 2s - 6 - \frac{1}{s+1} + \frac{8}{s+2} + \frac{9}{s+3}$$

Hence

$$\frac{X(s) - U(s)}{U(s)} = \frac{-1}{s+1} + \frac{2}{s+2} + \frac{1}{s+3}$$
$$\frac{sX(s) - (s+2)U(s)}{U(s)} = \frac{1}{s+1} - \frac{4}{s+2} - \frac{3}{s+3} \tag{4-87}$$
$$\frac{s^2X(s) - (s^2 + 2s - 6)U(s)}{U(s)} = \frac{-1}{s+1} + \frac{8}{s+2} + \frac{9}{s+3}$$

If we define

$$\begin{aligned} X_1(s) &= X(s) - U(s) \\ X_2(s) &= sX(s) - sU(s) - 2U(s) \\ X_3(s) &= s^2X(s) - s^2U(s) - 2sU(s) + 6U(s) \end{aligned} \tag{4-88}$$

Eq. (4–87) becomes

$$\begin{bmatrix} \frac{X_1(s)}{U(s)} \\ \frac{X_2(s)}{U(s)} \\ \frac{X_3(s)}{U(s)} \end{bmatrix} = \begin{bmatrix} -1 & 2 & 1 \\ 1 & -4 & -3 \\ -1 & 8 & 9 \end{bmatrix} \begin{bmatrix} \frac{1}{s+1} \\ \frac{1}{s+2} \\ \frac{1}{s+3} \end{bmatrix} \tag{4-89}$$

From Eq. (4–86) and Eq. (4–89), we obtain

$$\begin{bmatrix} X_1(s) \\ X_2(s) \\ X_3(s) \end{bmatrix} = \begin{bmatrix} -1 & 2 & 1 \\ 1 & -4 & -3 \\ -1 & 8 & 9 \end{bmatrix} \begin{bmatrix} Y_1(s) \\ Y_2(s) \\ Y_3(s) \end{bmatrix} \tag{4–90}$$

or

$$\begin{bmatrix} Y_1(s) \\ Y_2(s) \\ Y_3(s) \end{bmatrix} = \begin{bmatrix} -3 & -\frac{5}{2} & -\frac{1}{2} \\ -\frac{3}{2} & -2 & -\frac{1}{2} \\ 1 & \frac{3}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} X_1(s) \\ X_2(s) \\ X_3(s) \end{bmatrix} \tag{4–91}$$

In transforming the differential equation (4–84) into the transfer function, Eq. (4–85), we assumed the zero initial conditions. Again assuming the zero initial conditions, Eq. (4–88) can be written as

$$\begin{aligned} x_1 &= x - u \\ x_2 &= \dot{x} - \dot{u} - 2u \\ x_3 &= \ddot{x} - \ddot{u} - 2\dot{u} + 6u \end{aligned} \tag{4–92}$$

We hereafter relax zero initial conditions to arbitrary initial conditions. By use of Eq. (4–92), the initial condition terms $y_1(0)$, $y_2(0)$, and $y_3(0)$ can be written in terms of $x(0)$, $\dot{x}(0)$, $\ddot{x}(0)$ and $u(0)$, $\dot{u}(0)$, $\ddot{u}(0)$ as follows:

$$\begin{aligned} y_1(0) &= -\tfrac{1}{2}\ddot{x}(0) - \tfrac{5}{2}\dot{x}(0) - 3x(0) + \tfrac{1}{2}\ddot{u}(0) + \tfrac{7}{2}\dot{u}(0) + 5u(0) \\ y_2(0) &= -\tfrac{1}{2}\ddot{x}(0) - 2\dot{x}(0) - \tfrac{3}{2}x(0) + \tfrac{1}{2}\ddot{u}(0) + 3\dot{u}(0) + \tfrac{5}{2}u(0) \\ y_3(0) &= \tfrac{1}{2}\ddot{x}(0) + \tfrac{3}{2}\dot{x}(0) + x(0) - \tfrac{1}{2}\ddot{u}(0) - \tfrac{5}{2}\dot{u}(0) - u(0) \end{aligned}$$

The simplified state space representation of the system given by Eq. (4–84) or Eq. (4–85) is

$$\begin{aligned} \dot{y}_1 &= -y_1 + u \\ \dot{y}_2 &= -2y_2 + u \\ \dot{y}_3 &= -3y_3 + u \end{aligned}$$

The original variable x can be obtained in terms of y_1, y_2, and y_3 as follows:

$$\begin{aligned} x &= x_1 + u \\ &= -y_1 + 2y_2 + y_3 + u \end{aligned}$$

Since y_1, y_2, and y_3 can be easily obtained once u is specified, we can immediately find the response x.

The procedure just illustrated can be extended to the general case where $X(s)/U(s)$ is given by

$$\frac{X(s)}{U(s)} = \frac{b_0 s^n + b_1 s^{n-1} + \cdots + b_n}{s^n + a_1 s^{n-1} + \cdots + a_n}$$

The simplified state space representation of the system where the poles are distinct becomes

$$\begin{aligned} \frac{Y_1(s)}{U(s)} &= \frac{1}{s - \lambda_1} \\ \frac{Y_2(s)}{U(s)} &= \frac{1}{s - \lambda_2} \\ &\cdots \quad \cdots \\ \frac{Y_n(s)}{U(s)} &= \frac{1}{s - \lambda_n} \end{aligned}$$

If the system includes multiple poles of order m, then the state space representation of the system becomes that of Eq. (4–80).

The procedure for deriving an $n \times n$ transformation matrix $\mathbf{P}$, which relates $y_1, y_2, \ldots, y_n$ and $x_1, x_2, \ldots, x_n$, where the x_i are functions of x, u, and their derivatives, is identical to that just demonstrated by use of a simple example.

Effect of cancellation of common terms in transfer function If the transfer function is given as

$$\frac{X(s)}{U(s)} = \frac{K(s + a)(s + b)}{(s + a)(s + b)(s + c)(s + d)} \tag{4–93}$$

then cancellation of the common terms results in

$$\frac{\hat{X}(s)}{U(s)} = \frac{K}{(s + c)(s + d)} \tag{4–94}$$

The effect of cancellation of terms in the transfer function lowers the order of the system. We used $\hat{X}(s)$ instead of $X(s)$ in the canceled transfer function, since the order of $X(s)$ differs from that of $\hat{X}(s)$ by two. By the technique presented in this section we obtain a state space representation of Eq. (4–94) as follows:

$$\begin{aligned} \frac{Y_1(s)}{U(s)} &= \frac{1}{s + c} \\ \frac{Y_2(s)}{U(s)} &= \frac{1}{s + d} \end{aligned}$$

where $Y_1(s)$, $Y_2(s)$ and $\hat{X}(s)$ are related by

$$\begin{bmatrix} \hat{X}(s) \\ s\hat{X}(s) \end{bmatrix} = \begin{bmatrix} \dfrac{-K}{c - d} & \dfrac{K}{c - d} \\ \dfrac{cK}{c - d} & \dfrac{-dK}{c - d} \end{bmatrix} \begin{bmatrix} Y_1(s) \\ Y_2(s) \end{bmatrix}$$

Hence we obtain

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} -c & 0 \\ 0 & -d \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} [u] \tag{4–95}$$

Equation (4–95) is a state space representation of Eq. (4–94), but not that of

Eq. (4–93). To obtain a state space representation of Eq. (4–93), we must restore the order of the system. Let us define a state vector $\mathbf{z}$ as follows:

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ y_1 \\ y_2 \end{bmatrix}$$

Then a state space representation of Eq. (4–93) is given by

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \\ \dot{z}_4 \end{bmatrix} = \begin{bmatrix} -a & 0 & 0 & 0 \\ 0 & -b & 0 & 0 \\ 0 & 0 & -c & 0 \\ 0 & 0 & 0 & -d \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} [u]$$

4-5 STATE SPACE EQUATIONS OF SYSTEMS DESCRIBED BY DIFFERENCE EQUATIONS

Dynamic systems in which one or more variables can change only at discrete instants of time are called *discrete-time systems*. The behavior of discrete-time systems can be described in terms of difference equations. Note that in solving differential equations by use of a digital computer, we reduce differential equations to difference equations. By means of introducing discrete-time values, the problem of solving differential equations is modified into that of solving algebraic equations.

By use of a set of state variables, the behavior of discrete-time systems may be described conveniently in the form of vector matrix difference equations. In this section we shall first show that scalar difference equations can be written in the form of vector matrix difference equations and then show how vector matrix difference equations can be simplified.

Consider the following scalar difference equation:

$$x(k+n) + a_1 x(k+n-1) + \cdots + a_n x(k) = bu(k) \qquad (4\text{–}96)$$

where k denotes the kth sampling time or kth state in the evolution of the process; $x(k)$ is the system output at the kth sampling instant; and $u(k)$ is the input at the kth sampling instant. Let us define

$$\begin{aligned} x_1(k) &= x(k) \\ x_1(k+1) &= x_2(k) \\ x_2(k+1) &= x_3(k) \\ &\cdots \quad \cdots \\ x_{n-1}(k+1) &= x_n(k) \\ x_n(k+1) &= x(k+n) = -a_1 x_n(k) - a_2 x_{n-1}(k) - \cdots - a_n x_1(k) + bu(k) \end{aligned}$$

Then Eq. (4–96) can be written in the following form:

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ \cdot \\ \cdot \\ \cdot \\ x_{n-1}(k+1) \\ x_n(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & \cdots & 0 & 1 \\ -a_n & -a_{n-1} & \cdots & -a_2 & -a_1 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ \cdot \\ \cdot \\ \cdot \\ x_{n-1}(k) \\ x_n(k) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ b \end{bmatrix} [u(k)] \qquad (4\text{–}97)$$

or

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}u(k)$$

We shall next discuss simplification of vector matrix difference equations. The first matrix in the right-hand side of Eq. (4–97) can be simplified by means of the following transformation

$$\mathbf{x} = \mathbf{P}\mathbf{y}$$

or

$$\begin{bmatrix} x_1(k) \\ x_2(k) \\ \cdot \\ \cdot \\ \cdot \\ x_n(k) \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} \begin{bmatrix} y_1(k) \\ y_2(k) \\ \cdot \\ \cdot \\ \cdot \\ y_n(k) \end{bmatrix}$$

The matrix $\mathbf{P}$ is given by Eq. (3–37), if the roots of the following characteristic equation

$$\lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n = 0$$

are distinct. For the system with distinct eigenvalues, the simplified state space equation is given by

$$\begin{bmatrix} z_1(k+1) \\ z_2(k+1) \\ \cdot \\ \cdot \\ \cdot \\ z_n(k+1) \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} z_1(k) \\ z_2(k) \\ \cdot \\ \cdot \\ \cdot \\ z_n(k) \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix} [u(k)]$$

or

$$\mathbf{z}(k+1) = \mathbf{D}\mathbf{z}(k) + \mathbf{F}u(k)$$

where $\mathbf{z}(k)$ and $\mathbf{y}(k)$ are related by $\mathbf{y}(k) = \mathbf{V}\mathbf{z}(k)$ and $\mathbf{V}$ is defined by $\mathbf{P}^{-1}\mathbf{B} = \mathbf{V}\mathbf{F}$.

If the system has multiple eigenvalues, the state space equation must be modified in an obvious way.

Next, let us consider a state space representation of the following difference equation:

$$x(k+3) + a_1(k)x(k+2) + a_2(k)x(k+1) + a_3(k)x(k) = b_0(k)u(k+3) + b_1(k)u(k+2) + b_2(k)u(k+1) + b_3(k)u(k) \quad (4\text{–}98)$$

We shall show that Eq. (4–98) can be written in the following form:

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ x_3(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_3(k) & -a_2(k) & -a_1(k) \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \end{bmatrix} + \begin{bmatrix} c_1(k) \\ c_2(k) \\ c_3(k) \end{bmatrix} [u(k)] \quad (4\text{–}99)$$

Clearly, if we define

$$\begin{aligned} x_1(k) &= x(k) - c_0(k)u(k) \\ x_1(k+1) &= x_2(k) + c_1(k)u(k) \\ x_2(k+1) &= x_3(k) + c_2(k)u(k) \\ x_3(k+1) &= -a_3(k)x_1(k) - a_2(k)x_2(k) - a_1(k)x_3(k) + c_3(k)u(k) \end{aligned} \quad (4\text{–}100)$$

Eq. (4–98) can be transformed into Eq. (4–99). Then, we obtain

$$\begin{aligned} x(k+1) &= x_1(k+1) + c_0(k+1)u(k+1) \\ &= x_2(k) + c_1(k)u(k) + c_0(k+1)u(k+1) \\ x(k+2) &= x_2(k+1) + c_1(k+1)u(k+1) + c_0(k+2)u(k+2) \\ &= x_3(k) + c_2(k)u(k) + c_1(k+1)u(k+1) + c_0(k+2)u(k+2) \\ x(k+3) &= x_3(k+1) + c_2(k+1)u(k+1) + c_1(k+2)u(k+2) \\ &\quad + c_0(k+3)u(k+3) \\ &= -a_3(k)x_1(k) - a_2(k)x_2(k) - a_1(k)x_3(k) + c_3(k)u(k) \\ &\quad + c_2(k+1)u(k+1) + c_1(k+2)u(k+2) + c_0(k+3)u(k+3) \end{aligned}$$

Hence we obtain

$$\begin{aligned} &x(k+3) + a_1(k)x(k+2) + a_2(k)x(k+1) + a_3(k)x(k) \\ &= [-a_3(k)x_1(k) - a_2(k)x_2(k) - a_1(k)x_3(k) + c_3(k)u(k) \\ &\quad + c_2(k+1)u(k+1) + c_1(k+2)u(k+2) + c_0(k+3)u(k+3)] \\ &\quad + [a_1(k)x_3(k) + a_1(k)c_2(k)u(k) + a_1(k)c_1(k+1)u(k+1) \\ &\quad + a_1(k)c_0(k+2)u(k+2)] + [a_2(k)x_2(k) + a_2(k)c_1(k)u(k) \\ &\quad + a_2(k)c_0(k+1)u(k+1)] + [a_3(k)x_1(k) + a_3(k)c_0(k)u(k)] \end{aligned}$$

$$\begin{aligned} &= c_0(k+3)u(k+3) + [c_1(k+2) + a_1(k)c_0(k+2)]u(k+2) \\ &\quad + [c_2(k+1) + a_1(k)c_1(k+1) + a_2(k)c_0(k+1)]u(k+1) \\ &\quad + [c_3(k) + a_1(k)c_2(k) + a_2(k)c_1(k) + a_3(k)c_0(k)]u(k) \end{aligned}$$

The right-hand side of the preceding equation must be equal to the right-hand side of Eq. (4–98). Hence,

$$\begin{aligned} &[c_0(k+3) - b_0(k)]u(k+3) + [c_1(k+2) + a_1(k)c_0(k+2) - b_1(k)]u(k+2) \\ &\quad + [c_2(k+1) + a_1(k)c_1(k+1) + a_2(k)c_0(k+1) - b_2(k)]u(k+1) \\ &\quad + [c_3(k) + a_1(k)c_2(k) + a_2(k)c_1(k) + a_3(k)c_0(k) - b_3(k)]u(k) = 0 \end{aligned} \quad (4\text{–}101)$$

If we equate the coefficients of $u(k+3)$, $u(k+2)$, $u(k+1)$, and $u(k)$ in Eq. (4–101) to zero, respectively, then Eq. (4–98) is always satisfied. Therefore,

$$\begin{aligned} c_0(k+3) &= b_0(k) \\ c_1(k+2) + a_1(k)\,c_0(k+2) &= b_1(k) \\ c_2(k+1) + a_1(k)\,c_1(k+1) + a_2(k)\,c_0(k+1) &= b_2(k) \\ c_3(k) + a_1(k)\,c_2(k) + a_2(k)\,c_1(k) + a_3(k)\,c_0(k) &= b_3(k) \end{aligned}$$

it follows that

$$\begin{aligned} c_0(k) &= b_0(k-3) \\ c_1(k) &= b_1(k-2) - a_1(k-2)\,b_0(k-3) \\ c_2(k) &= b_2(k-1) - a_2(k-1)\,b_0(k-3) \\ &\quad - a_1(k-1)[b_1(k-2) - a_1(k-2)\,b_0(k-3)] \\ c_3(k) &= b_3(k) - a_3(k)\,b_0(k-3) \\ &\quad - [a_2(k) - a_1(k)\,a_1(k-1)]\,[b_1(k-2) - a_1(k-2)\,b_0(k-3)] \\ &\quad - a_1(k)\,[b_2(k-1) - a_2(k-1)\,b_0(k-3)] \end{aligned} \quad (4\text{–}102)$$

Equations (4–102) determine $c_0(k)$, $c_1(k)$, $c_2(k)$, and $c_3(k)$. The initial conditions are given by

$$\begin{aligned} x_1(0) &= x(0) - c_0(0)\,u(0) \\ x_2(0) &= x(1) - c_0(1)\,u(1) - c_1(0)\,u(0) \\ x_3(0) &= x(2) - c_0(2)\,u(2) - c_1(1)\,u(1) - c_2(0)\,u(0) \end{aligned}$$

We can extend the preceding technique to the nth-order difference equation. Consider

$$\begin{aligned} &x(k+n) + a_1(k)\,x(k+n-1) + \cdots + a_n(k)\,x(k) \\ &= b_0(k)\,u(k+n) + b_1(k)\,u(k+n-1) + \cdots + b_n(k)\,u(k) \\ &\qquad (k = 0, 1, 2, \ldots) \end{aligned} \quad (4\text{–}103)$$

If the initial data $x(0), x(1), \ldots, x(n-1)$ are known, Eq. (4–103) can be given in the following form:

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ \cdot \\ \cdot \\ \cdot \\ x_{n-1}(k+1) \\ x_n(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & 1 \\ -a_n(k) & -a_{n-1}(k) & \cdots & -a_1(k) \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ \cdot \\ \cdot \\ \cdot \\ x_{n-1}(k) \\ x_n(k) \end{bmatrix}$$

$$+ \begin{bmatrix} c_1(k) \\ c_2(k) \\ \cdot \\ \cdot \\ \cdot \\ c_{n-1}(k) \\ c_n(k) \end{bmatrix} [u(k)]$$

$$(k = 0, 1, 2, \ldots)$$

The $x_i(k+1)$ $(i = 1, 2, \ldots, n)$ are defined by

$$\begin{aligned} x(k) &= x_1(k) + c_0(k)\, u(k) \\ x_1(k+1) &= x_2(k) + c_1(k)\, u(k) \\ x_2(k+1) &= x_3(k) + c_2(k)\, u(k) \\ &\cdots \quad \cdots \quad \cdots \quad \cdots \\ x_{n-1}(k+1) &= x_n(k) + c_{n-1}(k)\, u(k) \\ x_n(k+1) &= -a_n(k)\, x_1(k) - a_{n-1}(k)\, x_2(k) - \cdots \\ &\quad -a_1(k)\, x_n(k) + c_n(k)\, u(k) \end{aligned}$$

where the $c_i(k)$ are determined from

$$\begin{aligned} c_0(k) &= b_0(k-n) \\ c_1(k) &= b_1(k-n+1) - a_1(k-n+1)\, c_0(k) \\ c_2(k) &= b_2(k-n+2) - a_2(k-n+2)\, c_0(k) - a_1(k-n+2)\, c_1(k) \\ &\cdots \quad \cdots \quad \cdots \\ c_n(k) &= b_n(k) - a_n(k)\, c_0(k) - \cdots - a_2(k)\, c_{n-2}(k) - a_1(k)\, c_{n-1}(k) \end{aligned}$$

The initial conditions are given by

$$\begin{aligned} x_1(0) &= x(0) - c_0(0)\, u(0) \\ x_2(0) &= x(1) - c_0(1)\, u(1) - c_1(0)\, u(0) \\ x_3(0) &= x(2) - c_0(2)\, u(2) - c_1(1)\, u(1) - c_2(0)\, u(0) \\ &\cdots \quad \cdots \quad \cdots \\ x_n(0) &= x(n-1) - c_0(n-1)\, u(n-1) - c_1(n-2)\, u(n-2) - \cdots \\ &\quad - c_{n-2}(1)\, u(1) - c_{n-1}(0)\, u(0) \end{aligned}$$

EXERCISE PROBLEMS AND SOLUTIONS

Problem A-4-1 Write the state space equation of the system shown in Fig. 4–5 in which x_1, x_2, and x_3 constitute the state vector. Simplify it to the form where the coefficient matrix of a new state vector is in the diagonal form and the elements of the coefficient matrix of the forcing function u are unity.

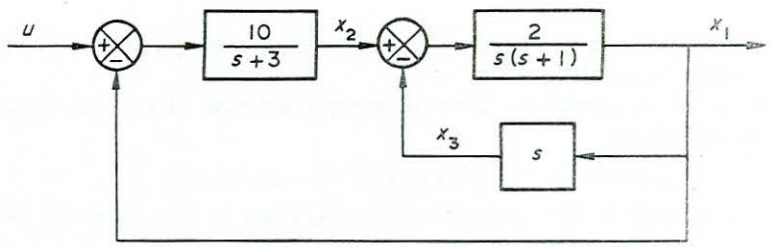


Figure 4-5 Control system.

Solution The state space equation of the system is obtained by writing the input output relationship for each transfer function and then taking the inverse Laplace transform. From the diagram we obtain

$$\begin{aligned} sX_1(s) &= X_3(s) \\ (s+3)X_2(s) &= 10[U(s) - X_1(s)] \\ s(s+1)X_1(s) &= 2[X_2(s) - X_3(s)] \end{aligned}$$

it follows that

$$\begin{aligned} \dot{x}_1 &= x_3 \\ \dot{x}_2 &= -10x_1 - 3x_2 + 10u \\ \dot{x}_3 &= 2x_2 - 3x_3 \end{aligned}$$

which can be rewritten as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ -10 & -3 & 0 \\ 0 & 2 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 10 \\ 0 \end{bmatrix} [u]$$

or

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u$$

The eigenvalues of $\mathbf{A}$ are

$$\lambda_1 = -5, \qquad \lambda_2 = -\frac{1}{2} - j\frac{\sqrt{15}}{2}, \qquad \lambda_3 = -\frac{1}{2} + j\frac{\sqrt{15}}{2}$$

A transformation matrix $\mathbf{P}$ for this system may be obtained as

$$\mathbf{P} = \begin{bmatrix} 1 & 1 & 1 \\ 5 & -\frac{5}{2} - j\frac{\sqrt{15}}{2} & -\frac{5}{2} + j\frac{\sqrt{15}}{2} \\ -5 & -\frac{1}{2} - j\frac{\sqrt{15}}{2} & -\frac{1}{2} + j\frac{\sqrt{15}}{2} \end{bmatrix}$$

If we put $\mathbf{x} = \mathbf{Py}$, then we obtain

$$\dot{\mathbf{y}} = \mathbf{P}^{-1}\mathbf{APy} + \mathbf{P}^{-1}\mathbf{B}u$$

The inverse of $\mathbf{P}$ is found as

$$\mathbf{P}^{-1} = \begin{bmatrix} \frac{1}{6} & \frac{1}{12} & -\frac{1}{12} \\ \frac{5}{12} + j\frac{\sqrt{15}}{12} & -\frac{1}{24} + j\frac{\sqrt{15}}{40} & \frac{1}{24} + j\frac{\sqrt{15}}{24} \\ \frac{5}{12} - j\frac{\sqrt{15}}{12} & -\frac{1}{24} - j\frac{\sqrt{15}}{40} & \frac{1}{24} - j\frac{\sqrt{15}}{24} \end{bmatrix}$$

Hence

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} -5 & 0 & 0 \\ 0 & -\frac{1}{2} - j\frac{\sqrt{15}}{2} & 0 \\ 0 & 0 & -\frac{1}{2} + j\frac{\sqrt{15}}{2} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + \begin{bmatrix} \frac{5}{6} \\ -\frac{5}{12} + j\frac{\sqrt{15}}{4} \\ -\frac{5}{12} - j\frac{\sqrt{15}}{4} \end{bmatrix} [u]$$

Put $\mathbf{y} = \mathbf{Vz}$ where $\mathbf{V}$ is defined by

$$\mathbf{V} = \begin{bmatrix} \frac{5}{6} & 0 & 0 \\ 0 & -\frac{5}{12} + j\frac{\sqrt{15}}{4} & 0 \\ 0 & 0 & -\frac{5}{12} - j\frac{\sqrt{15}}{4} \end{bmatrix}$$

Then, the state space equation of the system can be simplified into

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \end{bmatrix} = \begin{bmatrix} -5 & 0 & 0 \\ 0 & -\frac{1}{2} - j\frac{\sqrt{15}}{2} & 0 \\ 0 & 0 & -\frac{1}{2} + j\frac{\sqrt{15}}{2} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} [u]$$

Clearly, $\mathbf{z}$ is related to $\mathbf{x}$ by

$$\mathbf{x} = \mathbf{PVz}$$

Problem A-4-2 Obtain a state space equation of the following system:

$$\dddot{x} + 4\ddot{x} + 5\dot{x} + 2x = 5u$$

Solution The characteristic equation has the following roots:

$$\lambda_1 = -1, \quad \lambda_2 = -1, \quad \lambda_3 = -2$$

The following transformation matrix

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & -2 \\ 1 & -2 & 4 \end{bmatrix}$$

will transform the given differential equation into the following form:

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + \begin{bmatrix} -5 \\ 5 \\ 5 \end{bmatrix} [u] \qquad (4\text{–}104)$$

or $\dot{\mathbf{y}} = \mathbf{Jy} + \mathbf{P}^{-1}\mathbf{B}u$. To simplify Eq. (4–104) further, let

$$\mathbf{y} = \mathbf{Vz}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} 5 & -5 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

Since

$$\mathbf{V}^{-1} = \begin{bmatrix} \frac{1}{5} & \frac{1}{5} & 0 \\ 0 & \frac{1}{5} & 0 \\ 0 & 0 & \frac{1}{5} \end{bmatrix}$$

$\mathbf{V}^{-1}\mathbf{P}^{-1}\mathbf{B}u$ and $\mathbf{V}^{-1}\mathbf{JV}$ are obtained as follows:

$$\mathbf{V}^{-1}\mathbf{P}^{-1}\mathbf{B}u = \begin{bmatrix} \frac{1}{5} & \frac{1}{5} & 0 \\ 0 & \frac{1}{5} & 0 \\ 0 & 0 & \frac{1}{5} \end{bmatrix} \begin{bmatrix} -5 \\ 5 \\ 5 \end{bmatrix} [u] = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} [u]$$

$$\mathbf{V}^{-1}\mathbf{JV} = \begin{bmatrix} \frac{1}{5} & \frac{1}{5} & 0 \\ 0 & \frac{1}{5} & 0 \\ 0 & 0 & \frac{1}{5} \end{bmatrix} \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} 5 & -5 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix}$$

The given differential equation is then transformed into the following form:

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} [u]$$

Since $\mathbf{x} = \mathbf{Py} = \mathbf{PVz}$, the vector $\mathbf{x}$ can be transformed into $\mathbf{z}$ directly by the following transformation matrix:

$$\mathbf{PV} = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & -2 \\ 1 & -2 & 4 \end{bmatrix} \begin{bmatrix} 5 & -5 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{bmatrix} = \begin{bmatrix} 5 & -5 & 5 \\ -5 & 10 & -10 \\ 5 & -15 & 20 \end{bmatrix}$$

[Notice that $\mathbf{PV}$ is the same as $\mathbf{P}_p$ given by Eq. (4–82).]

Problem A-4-3 Obtain a state space equation of the following system:

$$\dddot{x} + 6\ddot{x} + 11\dot{x} + 6 = 3\ddot{u} + 2\dot{u} + u$$

Solution Let us use Method 2 of Section 4–2. The given differential equation can be written as

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix}=\begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ -6 & -11 & -6\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}+\begin{bmatrix}c_1\\ c_2\\ c_3\end{bmatrix}[u]$$

where c_1, c_2, and c_3 are determined from Eq. (4–37) as follows:

$$c_1 = 3, \qquad c_2 = -16, \qquad c_3 = 64$$

Since the characteristic equation has roots -1, -2, and -3 the preceding state space equation can be written as follows:

$$\begin{bmatrix}\dot{y}_1\\ \dot{y}_2\\ \dot{y}_3\end{bmatrix}=\begin{bmatrix}-1 & 0 & 0\\ 0 & -2 & 0\\ 0 & 0 & -3\end{bmatrix}\begin{bmatrix}y_1\\ y_2\\ y_3\end{bmatrix}+\begin{bmatrix}1 & 1 & 1\\ -1 & -2 & -3\\ 1 & 4 & 9\end{bmatrix}^{-1}\begin{bmatrix}3\\ -16\\ 64\end{bmatrix}[u]$$

$$=\begin{bmatrix}-1 & 0 & 0\\ 0 & -2 & 0\\ 0 & 0 & -3\end{bmatrix}\begin{bmatrix}y_1\\ y_2\\ y_3\end{bmatrix}+\begin{bmatrix}1\\ -9\\ 11\end{bmatrix}[u]$$

By use of a second transformation

$$\mathbf{y} = \mathbf{V}\mathbf{z}$$

where

$$\mathbf{V}=\begin{bmatrix}1 & 0 & 0\\ 0 & -9 & 0\\ 0 & 0 & 11\end{bmatrix}$$

we obtain

$$\begin{bmatrix}\dot{z}_1\\ \dot{z}_2\\ \dot{z}_3\end{bmatrix}=\begin{bmatrix}-1 & 0 & 0\\ 0 & -2 & 0\\ 0 & 0 & -3\end{bmatrix}\begin{bmatrix}z_1\\ z_2\\ z_3\end{bmatrix}+\begin{bmatrix}1\\ 1\\ 1\end{bmatrix}[u]$$

Vector $\mathbf{x}$ and vector $\mathbf{z}$ are related by

$$\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}=\begin{bmatrix}1 & 1 & 1\\ -1 & -2 & -3\\ 1 & 4 & 9\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ 0 & -9 & 0\\ 0 & 0 & 11\end{bmatrix}\begin{bmatrix}z_1\\ z_2\\ z_3\end{bmatrix}$$

The original variable x is related to x_1 as follows:

$$x = x_1$$

The initial conditions on $\mathbf{z}$ are obtained as

$$\begin{bmatrix}z_1(0)\\ z_2(0)\\ z_3(0)\end{bmatrix}=\begin{bmatrix}3 & \frac{5}{2} & \frac{1}{2}\\ \frac{1}{3} & \frac{4}{9} & \frac{1}{9}\\ \frac{1}{11} & \frac{3}{22} & \frac{1}{22}\end{bmatrix}\begin{bmatrix}x_1(0)\\ x_2(0)\\ x_3(0)\end{bmatrix}$$

and

$$x_1(0) = x(0)$$
$$x_2(0) = \dot{x}(0) - 3u(0)$$
$$x_3(0) = \ddot{x}(0) - 3\dot{u}(0) + 16u(0)$$

Problem A-4-4 Given the following matrix $\mathbf{P}_1$

$$\mathbf{P}_1=\begin{bmatrix}1 & -L & 1\\ \lambda_1 & 1-\lambda_1 L & \lambda_3\\ \lambda_1^2 & 2\lambda_1-\lambda_1^2 L & \lambda_3^2\end{bmatrix}$$

$$=\begin{bmatrix}1 & 0 & 1\\ \lambda_1 & 1 & \lambda_3\\ \lambda_1^2 & 2\lambda_1 & \lambda_3^2\end{bmatrix}+\begin{bmatrix}0 & -L & 0\\ 0 & -\lambda_1 L & 0\\ 0 & -\lambda_1^2 L & 0\end{bmatrix}$$

$$=\mathbf{P}+\begin{bmatrix}0 & -L & 0\\ 0 & -\lambda_1 L & 0\\ 0 & -\lambda_1^2 L & 0\end{bmatrix}$$

where L is a constant, show that the inverse of $\mathbf{P}_1$ is

$$\mathbf{P}_1^{-1}=\mathbf{P}^{-1}+\begin{bmatrix}\frac{L\Delta_1}{|\mathbf{P}|} & \frac{L\Delta_2}{|\mathbf{P}|} & \frac{L\Delta_3}{|\mathbf{P}|}\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

where Δ_1, Δ_2, Δ_3 are the cofactors of the elements of the second column of $\mathbf{P}$.

Solution The determinant of $\mathbf{P}_1$ and that of $\mathbf{P}$ are found as

$$|\mathbf{P}_1| = (\lambda_1 - \lambda_3)^2, \qquad |\mathbf{P}| = (\lambda_1 - \lambda_3)^2$$

Hence

$$|\mathbf{P}_1| = |\mathbf{P}|$$

The inverse of $\mathbf{P}_1$ is

$$\mathbf{P}_1^{-1}=\frac{1}{(\lambda_1-\lambda_3)^2}\begin{bmatrix}L\lambda_1\lambda_3(\lambda_1-\lambda_3)-2\lambda_1\lambda_3+\lambda_3^2 & L(\lambda_3^2-\lambda_1^2)+2\lambda_1 & L(\lambda_1-\lambda_3)-1\\ \lambda_1\lambda_3(\lambda_1-\lambda_3) & \lambda_3^2-\lambda_1^2 & \lambda_1-\lambda_3\\ \lambda_1^2 & -2\lambda_1 & 1\end{bmatrix}$$

Notice that the inverse of $\mathbf{P}$ can be obtained by letting $L = 0$ in $\mathbf{P}_1^{-1}$. Hence,

$$\mathbf{P}_1^{-1}-\mathbf{P}^{-1}=\frac{1}{(\lambda_1-\lambda_3)^2}\begin{bmatrix}L\lambda_1\lambda_3(\lambda_1-\lambda_3) & L(\lambda_3^2-\lambda_1^2) & L(\lambda_1-\lambda_3)\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

Notice also that

$$\lambda_1\lambda_3(\lambda_1-\lambda_3) = \Delta_1$$
$$\lambda_3^2-\lambda_1^2 = \Delta_2$$
$$\lambda_1-\lambda_3 = \Delta_3$$

Therefore, we obtain

$$\mathbf{P}_1^{-1} = \mathbf{P}^{-1} + \begin{bmatrix} \frac{L\Delta_1}{|\mathbf{P}|} & \frac{L\Delta_2}{|\mathbf{P}|} & \frac{L\Delta_3}{|\mathbf{P}|} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Problem A-4-5 Referring to Eq. (4–78), (1) prove that

$$\begin{bmatrix} \frac{L\Delta_1}{|\mathbf{P}|} & \frac{L\Delta_2}{|\mathbf{P}|} & \frac{L\Delta_3}{|\mathbf{P}|} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_3 & -a_2 & -a_1 \end{bmatrix} \begin{bmatrix} 0 & -L & 0 \\ 0 & -\lambda_1 L & 0 \\ 0 & -\lambda_1^2 L & 0 \end{bmatrix} = \mathbf{0}$$

where

$$L = \frac{1}{\lambda_1 - \lambda_3} = \text{constant}$$

$$a_1 = -(2\lambda_1 + \lambda_3)$$

$$a_2 = 2\lambda_1\lambda_3 + \lambda_1^2$$

$$a_3 = -\lambda_1^2\lambda_3$$

$$\mathbb{P} = \begin{bmatrix} 1 & 0 & 1 \\ \lambda_1 & 1 & \lambda_3 \\ \lambda_1^2 & 2\lambda_1 & \lambda_3^2 \end{bmatrix}$$

and Δ_1, Δ_2, Δ_3 are the cofactors of the elements of the second column of $\mathbf{P}$. (2) Prove also that

$$\begin{bmatrix} \frac{L\Delta_1}{|\mathbf{P}|} & \frac{L\Delta_2}{|\mathbf{P}|} & \frac{L\Delta_3}{|\mathbf{P}|} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_3 & -a_2 & -a_1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ \lambda_1 & 1 & \lambda_3 \\ \lambda_1^2 & 2\lambda_1 & \lambda_3^2 \end{bmatrix}$$
$$+ \begin{bmatrix} 1 & 0 & 1 \\ \lambda_1 & 1 & \lambda_3 \\ \lambda_1^2 & 2\lambda_1 & \lambda_3^2 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_3 & -a_2 & -a_1 \end{bmatrix} \begin{bmatrix} 0 & -L & 0 \\ 0 & -\lambda_1 L & 0 \\ 0 & -\lambda_1^2 L & 0 \end{bmatrix} = \mathbf{0}$$

Solution Noting that $a_3 + a_2\lambda_1 + a_1\lambda_1^2 = -\lambda_1^3$, we obtain

$$\begin{bmatrix} \frac{L\Delta_1}{|\mathbf{P}|} & \frac{L\Delta_2}{|\mathbf{P}|} & \frac{L\Delta_3}{|\mathbf{P}|} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_3 & -a_2 & -a_1 \end{bmatrix} \begin{bmatrix} 0 & -L & 0 \\ 0 & -\lambda_1 L & 0 \\ 0 & -\lambda_1^2 L & 0 \end{bmatrix}$$
$$= \begin{bmatrix} 0 & \frac{-L^2}{|\mathbf{P}|}(\Delta_1\lambda_1 + \Delta_2\lambda_1^2 + \Delta_3\lambda_1^3) & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Since

$$\Delta_1 = \lambda_1^2\lambda_3 - \lambda_1\lambda_3^2$$
$$\Delta_2 = \lambda_3^2 - \lambda_1^2$$
$$\Delta_3 = \lambda_1 - \lambda_3$$

we obtain

$$\Delta_1\lambda_1 + \Delta_2\lambda_1^2 + \Delta_3\lambda_1^3 = 0$$

This proves (1) of the problem.

Noting that

$$a_2 + 2a_1\lambda_1 = -3\lambda_1^2$$
$$a_3 + a_2\lambda_1 + a_1\lambda_1^2 = -\lambda_1^3$$
$$a_3 + a_2\lambda_3 + a_1\lambda_3^2 = -\lambda_3^3$$

we obtain

$$\begin{bmatrix} \frac{L\Delta_1}{|\mathbf{P}|} & \frac{L\Delta_2}{|\mathbf{P}|} & \frac{L\Delta_3}{|\mathbf{P}|} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_3 & -a_2 & -a_1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ \lambda_1 & 1 & \lambda_3 \\ \lambda_1^2 & 2\lambda_1 & \lambda_3^2 \end{bmatrix}$$
$$+ \begin{bmatrix} 1 & 0 & 1 \\ \lambda_1 & 1 & \lambda_3 \\ \lambda_1^2 & 2\lambda_1 & \lambda_3^2 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_3 & -a_2 & -a_1 \end{bmatrix} \begin{bmatrix} 0 & -L & 0 \\ 0 & -\lambda_1 L & 0 \\ 0 & -\lambda_1^2 L & 0 \end{bmatrix}$$
$$= \frac{L}{|\mathbf{P}|} \begin{bmatrix} \Delta_1\lambda_1 + \Delta_2\lambda_1^2 + \Delta_3\lambda_1^3 & \Delta_1 + 2\Delta_2\lambda_1 + 3\Delta_3\lambda_1^2 & \Delta_1\lambda_3 + \Delta_2\lambda_3^2 + \Delta_3\lambda_3^3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$
$$+ \frac{1}{|\mathbf{P}|} \begin{bmatrix} \lambda_3^2 - 2\lambda_1\lambda_3 & 2\lambda_1 & -1 \\ \lambda_1^2\lambda_3 - \lambda_1\lambda_3^2 & \lambda_3^2 - \lambda_1^2 & \lambda_1 - \lambda_3 \\ \lambda_1^2 & -2\lambda_1 & 1 \end{bmatrix} \begin{bmatrix} 0 & -\lambda_1 L & 0 \\ 0 & -\lambda_1^2 L & 0 \\ 0 & -\lambda_1^3 L & 0 \end{bmatrix}$$
$$= \frac{L}{|\mathbf{P}|} \begin{bmatrix} 0 & \Delta_1 + 2\Delta_2\lambda_1 + 3\Delta_3\lambda_1^2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \frac{L}{|\mathbf{P}|} \begin{bmatrix} 0 & -\lambda_1^3 - \lambda_1\lambda_3^2 + 2\lambda_1^2\lambda_3 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$
$$= \frac{L}{|\mathbf{P}|} \begin{bmatrix} 0 & \Delta_1 + 2\Delta_2\lambda_1 + 3\Delta_3\lambda_1^2 - \lambda_1^3 - \lambda_1\lambda_3^2 + 2\lambda_1^2\lambda_3 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Notice that

$$\Delta_1 + 2\Delta_2\lambda_1 + 3\Delta_3\lambda_1^2 - \lambda_1^3 - \lambda_1\lambda_3^2 + 2\lambda_1^2\lambda_3$$
$$= \lambda_1^2\lambda_3 - \lambda_1\lambda_3^2 + 2(\lambda_3^2 - \lambda_1^2)\lambda_1 + 3(\lambda_1 - \lambda_3)\lambda_1^2 - \lambda_1^3 - \lambda_1\lambda_3^2 + 2\lambda_1^2\lambda_3$$
$$= 0$$

This proves (2) of the problem.

Problem A-4-6 Given the following state space equation:

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix}=\begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ 1 & -3 & 3\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}+\begin{bmatrix}1\\ 1\\ 1\end{bmatrix}[u]$$

obtain the corresponding scalar differential equation in terms of x_1.

Solution The characteristic equation is

$$\lambda^3 - 3\lambda^2 + 3\lambda - 1 = 0$$

The desired scalar differential equation can then be written as

$$\dddot{x}_1 - 3\ddot{x}_1 + 3\dot{x}_1 - x_1 = b_0\ddot{u} + b_1\ddot{u} + b_2\dot{u} + b_3u$$

The constants b_0, b_1, b_2, and b_3 are found from Eq. (4–46) as follows:

$$b_0 = 0, \qquad b_1 = 1, \qquad b_2 = -2, \qquad b_3 = 1$$

Hence we obtain

$$\dddot{x}_1 - 3\ddot{x}_1 + 3\dot{x}_1 - x_1 = \ddot{u} - 2\dot{u} + u$$

or

$$(p-1)^3x_1 = (p-1)^2u, \qquad p = \frac{d}{dt}$$

Problem A-4-7 Consider a nonsingular matrix $\mathbf{P}$ consisting of linearly independent column vectors $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_n$, or

$$\mathbf{P} = [\mathbf{P}_1 \mid \mathbf{P}_2 \mid \cdots \mid \mathbf{P}_n]$$

For an $n \times n$ matrix $\mathbf{A}$ if $\mathbf{AP}_i$ can be expressed as

$$\mathbf{AP}_i = b_{1i}\mathbf{P}_1 + b_{2i}\mathbf{P}_2 + \cdots + b_{ni}\mathbf{P}_n$$

where the b_{ji} $(i, j = 1, 2, \ldots, n)$ are constants, then show that

$$\mathbf{P}^{-1}\mathbf{AP} = \begin{bmatrix}b_{11} & b_{12} & \cdots & b_{1n}\\ b_{21} & b_{22} & \cdots & b_{2n}\\ \vdots & \vdots & & \vdots\\ b_{n1} & b_{n2} & \cdots & b_{nn}\end{bmatrix} = \mathbf{B}$$

Solution Notice that $\mathbf{AP}_i$ is the ith column of $\mathbf{AP}$. Hence

$$\mathbf{AP} = [\mathbf{AP}_1 \mid \mathbf{AP}_2 \mid \cdots \mid \mathbf{AP}_n] = \begin{bmatrix}\sum_{j=1}^{n} b_{j1}p_{1j} & \cdots & \sum_{j=1}^{n} b_{jn}p_{1j}\\ \sum_{j=1}^{n} b_{j1}p_{2j} & \cdots & \sum_{j=1}^{n} b_{jn}p_{2j}\\ \vdots & & \vdots\\ \sum_{j=1}^{n} b_{j1}p_{nj} & \cdots & \sum_{j=1}^{n} b_{jn}p_{nj}\end{bmatrix}$$

$$= \begin{bmatrix}p_{11} & p_{12} & \cdots & p_{1n}\\ p_{21} & p_{22} & \cdots & p_{2n}\\ \vdots & \vdots & & \vdots\\ p_{n1} & p_{n2} & \cdots & p_{nn}\end{bmatrix}\begin{bmatrix}b_{11} & b_{12} & \cdots & b_{1n}\\ b_{21} & b_{22} & \cdots & b_{2n}\\ \vdots & \vdots & & \vdots\\ b_{n1} & b_{n2} & \cdots & b_{nn}\end{bmatrix} = \mathbf{PB}$$

Since $\mathbf{P}$ is nonsingular, we obtain

$$\mathbf{P}^{-1}\mathbf{AP} = \mathbf{B}$$

From this we see that if we denote by $\mathbf{Q}_i$ the ith row vector of $\mathbf{P}^{-1}$, then

$$\mathbf{Q}_i\mathbf{AP}_j = b_{ij}$$

Problem A-4-8 Suppose that $\mathbf{A}$ is an $n \times n$ nonsingular matrix and $\mathbf{B}$ is an $n \times 1$ matrix. Suppose also that $\mathbf{B}, \mathbf{AB}, \mathbf{A}^2\mathbf{B}, \ldots, \mathbf{A}^{n-1}\mathbf{B}$ are linearly independent. Define an $n \times n$ nonsingular matrix $\mathbf{P}$ such that its columns consist of $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_n$, or

$$\mathbf{P} = [\mathbf{P}_1 \mid \mathbf{P}_2 \mid \cdots \mid \mathbf{P}_n]$$

where

$$\mathbf{P}_1 = \mathbf{B}$$
$$\mathbf{P}_2 = \mathbf{AB}$$
$$\cdots$$
$$\mathbf{P}_n = \mathbf{A}^{n-1}\mathbf{B}$$

Show that it is possible to write $\mathbf{P}^{-1}\mathbf{AP}$ as follows:

$$\mathbf{P}^{-1}\mathbf{AP} = \begin{bmatrix}0 & 0 & \cdots & 0 & -a_n\\ 1 & 0 & \cdots & 0 & -a_{n-1}\\ 0 & 1 & \cdots & 0 & -a_{n-2}\\ \vdots & \vdots & & \vdots & \vdots\\ 0 & 0 & \cdots & 1 & -a_1\end{bmatrix} \equiv \mathbf{C}$$

Solution The matrix $\mathbf{P}$ is given by

$$\mathbf{P} = [\mathbf{P}_1 \mid \mathbf{P}_2 \mid \cdots \mid \mathbf{P}_n] = [\mathbf{B} \mid \mathbf{AB} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{B}]$$

First, let us compute $\mathbf{AP}$.

$$\mathbf{AP} = \mathbf{A}[\mathbf{B} \mid \mathbf{AB} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{B}] = [\mathbf{AB} \mid \mathbf{A}^2\mathbf{B} \mid \cdots \mid \mathbf{A}^n\mathbf{B}]$$

Then compute $\mathbf{PC}$.

$$\mathbf{PC} = [\mathbf{B} \mid \mathbf{AB} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{B}]\begin{bmatrix}0 & 0 & \cdots & 0 & -a_n\\ 1 & 0 & \cdots & 0 & -a_{n-1}\\ 0 & 1 & \cdots & 0 & -a_{n-2}\\ \vdots & \vdots & & \vdots & \vdots\\ 0 & 0 & \cdots & 1 & -a_1\end{bmatrix}$$

$$= [\mathbf{AB} \mid \mathbf{A}^2\mathbf{B} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{B} \mid -a_n\mathbf{B} - a_{n-1}\mathbf{AB} - \cdots - a_1\mathbf{A}^{n-1}\mathbf{B}]$$

Since the column vectors $\mathbf{B}, \mathbf{AB}, \ldots, \mathbf{A}^{n-1}\mathbf{B}$ are linearly independent, all other n vectors can be formed from a linear combination of these independent vectors. If we pick $a_1, a_2, \ldots, a_n$ such that

$$\mathbf{A}^n\mathbf{B} = -a_n\mathbf{B} - a_{n-1}\mathbf{AB} - \cdots - a_1\mathbf{A}^{n-1}\mathbf{B}$$

then

$$\mathbf{AP} = \mathbf{PC}$$

Since P is nonsingular, we obtain

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{C} = \begin{bmatrix} 0 & 0 & \cdots & 0 & -a_n \\ 1 & 0 & \cdots & 0 & -a_{n-1} \\ 0 & 1 & \cdots & 0 & -a_{n-2} \\ \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -a_1 \end{bmatrix}$$

Problem A-4-9 Consider the system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u \tag{4-105}$$

Show that if A has rank n and column vectors $\mathbf{B}, \mathbf{AB}, \mathbf{A}^2\mathbf{B}, \ldots, \mathbf{A}^{n-1}\mathbf{B}$ are linearly independent, then by means of a suitable transformation, Eq. (4-105) can be transformed into the following form:

$$\dot{\mathbf{y}} = \mathbf{C}\mathbf{y} + \mathbf{F}u \tag{4-106}$$

where

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & \cdots & 0 & -a_n \\ 1 & 0 & \cdots & 0 & -a_{n-1} \\ 0 & 1 & \cdots & 0 & -a_{n-2} \\ \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -a_1 \end{bmatrix}, \qquad \mathbf{F} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}$$

Here, $a_1, a_2, \ldots, a_n$ are constants determined by A and B. (The elements of A are a_{ij} which are different from $a_1, a_2, \ldots, a_n$.)

Show also that by eliminating $y_1, y_2, \ldots, y_{n-1}$ from Eq. (4-106), we can obtain the following scalar differential equation:

$$\overset{(n)}{y_n} + a_1\overset{(n-1)}{y_n} + a_2\overset{(n-2)}{y_n} + \cdots + a_{n-1}\dot{y}_n + a_n y_n = u$$

Solution Define the matrix H to consist of n linearly independent column vectors $\mathbf{B}, \mathbf{AB}, \mathbf{A}^2\mathbf{B}, \ldots, \mathbf{A}^{n-1}\mathbf{B}$, or

$$\mathbf{H} = [\mathbf{B} \mid \mathbf{AB} \mid \mathbf{A}^2\mathbf{B} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{B}]$$

Then H is nonsingular and $\mathbf{H}^{-1}$ exists. Let us put $\mathbf{x} = \mathbf{H}\mathbf{y}$. Then Eq. (4-105) becomes

$$\dot{\mathbf{y}} = \mathbf{H}^{-1}\mathbf{A}\mathbf{H}\mathbf{y} + \mathbf{H}^{-1}\mathbf{B}u$$

Observe that

$$\mathbf{H}\begin{bmatrix} 1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix} = [\mathbf{B} \mid \mathbf{AB} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{B}]\begin{bmatrix} 1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix} = \mathbf{B}$$

Thus

$$\mathbf{H}^{-1}\mathbf{B} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}$$

Computing AH we obtain

$$\begin{aligned} \mathbf{AH} &= \mathbf{A}[\mathbf{B} \mid \mathbf{AB} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{B}] \\ &= [\mathbf{AB} \mid \mathbf{A}^2\mathbf{B} \mid \cdots \mid \mathbf{A}^n\mathbf{B}] \end{aligned} \tag{4-107}$$

Next compute HC.

$$\mathbf{HC} = \mathbf{H}\begin{bmatrix} 0 & 0 & \cdots & 0 & -a_n \\ 1 & 0 & \cdots & 0 & -a_{n-1} \\ 0 & 1 & \cdots & 0 & -a_{n-2} \\ \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -a_1 \end{bmatrix}$$

$$= [\mathbf{B} \mid \mathbf{AB} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{B}]\begin{bmatrix} 0 & 0 & \cdots & 0 & -a_n \\ 1 & 0 & \cdots & 0 & -a_{n-1} \\ 0 & 1 & \cdots & 0 & -a_{n-2} \\ \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -a_1 \end{bmatrix}$$

$$= [\mathbf{AB} \mid \mathbf{A}^2\mathbf{B} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{B} \mid -a_n\mathbf{B} - a_{n-1}\mathbf{AB} - \cdots - a_1\mathbf{A}^{n-1}\mathbf{B}] \tag{4-108}$$

Comparing the right-hand side of Eq. (4-107) and the right-hand side of Eq. (4-108), it is obvious that the first $n-1$ corresponding columns are equal. If the a_i are chosen such that

$$\mathbf{A}^n\mathbf{B} = -a_n\mathbf{B} - a_{n-1}\mathbf{AB} - \cdots - a_1\mathbf{A}^{n-1}\mathbf{B}$$

which is equivalent to

$$\mathbf{A}^n\mathbf{B} = [\mathbf{B} \mid \mathbf{AB} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{B}]\begin{bmatrix} -a_n \\ -a_{n-1} \\ \vdots \\ \vdots \\ -a_1 \end{bmatrix} = \mathbf{H}\begin{bmatrix} -a_n \\ -a_{n-1} \\ \vdots \\ \vdots \\ -a_1 \end{bmatrix}$$

or

$$\mathbf{H}^{-1}\mathbf{A}^n\mathbf{B} = \begin{bmatrix} -a_n \\ -a_{n-1} \\ \vdots \\ \vdots \\ -a_1 \end{bmatrix} \tag{4-109}$$

then the right-hand side of Eq. (4–107) and that of Eq. (4–108) are equal. This means that if we determine the a_i according to Eq. (4–109), then Eq. (4–105) can be transformed into Eq. (4–106).

Since Eq. (4–106) is equivalent to

$$\dot{y}_1 = -a_n y_n + u$$
$$\dot{y}_2 = y_1 - a_{n-1} y_n$$
$$\cdots$$
$$\dot{y}_n = y_{n-1} - a_1 y_n$$

by eliminating $y_1, y_2, \ldots, y_{n-1}$ we obtain

$$y_n^{(n)} + a_1 y_n^{(n-1)} + a_2 y_n^{(n-2)} + \cdots + a_{n-1}\dot{y}_n + a_n y_n = u$$

Problem A-4-10 Show that, by a suitable transformation, the system

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix} = \begin{bmatrix}0 & 1 & 0\\ 3 & 0 & 2\\ -12 & -7 & -6\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix} + \begin{bmatrix}-1\\ 2\\ 3\end{bmatrix}[u]$$

can be transformed into

$$\begin{bmatrix}\dot{z}_1\\ \dot{z}_2\\ \dot{z}_3\end{bmatrix} = \begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ -6 & -11 & -6\end{bmatrix}\begin{bmatrix}z_1\\ z_2\\ z_3\end{bmatrix} + \begin{bmatrix}0\\ 0\\ 1\end{bmatrix}[u]$$

Solution Let us perform the transformation

$$\mathbf{x} = \mathbf{H}\mathbf{y}$$

where

$$\mathbf{H} = [\mathbf{B}\,|\,\mathbf{AB}\,|\,\mathbf{A}^2\mathbf{B}]$$

and

$$\mathbf{A} = \begin{bmatrix}0 & 1 & 0\\ 3 & 0 & 2\\ -12 & -7 & -6\end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix}-1\\ 2\\ 3\end{bmatrix}$$

After a simple computation, we obtain

$$\mathbf{H} = \begin{bmatrix}-1 & 2 & 3\\ 2 & 3 & -34\\ 3 & -20 & 75\end{bmatrix}$$

and

$$\mathbf{H}^{-1} = \frac{-1}{196}\begin{bmatrix}-455 & -210 & -77\\ -252 & -84 & -28\\ -49 & -14 & -7\end{bmatrix}$$

Since

$$\mathbf{A}^3\mathbf{B} = \begin{bmatrix}-34\\ 159\\ -248\end{bmatrix}$$

we obtain by use of Eq. (4–60)

$$-a_3 = \frac{-1}{196}[(-455)(-34) + (-210)(159) + (-77)(-248)] = -6$$

$$-a_2 = \frac{-1}{196}[(-252)(-34) + (-84)(159) + (-28)(-248)] = -11$$

$$-a_1 = \frac{-1}{196}[(-49)(-34) + (-14)(159) + (-7)(-248)] = -6$$

Hence

$$\dot{\mathbf{y}} = \mathbf{H}^{-1}\mathbf{A}\mathbf{H}\mathbf{y} + \mathbf{H}^{-1}\mathbf{B}u$$

becomes

$$\begin{bmatrix}\dot{y}_1\\ \dot{y}_2\\ \dot{y}_3\end{bmatrix} = \begin{bmatrix}0 & 0 & -6\\ 1 & 0 & -11\\ 0 & 1 & -6\end{bmatrix}\begin{bmatrix}y_1\\ y_2\\ y_3\end{bmatrix} + \begin{bmatrix}1\\ 0\\ 0\end{bmatrix}[u]$$

By eliminating y_1 and y_2 from the foregoing equation, we obtain

$$\dddot{y}_3 + 6\ddot{y}_3 + 11\dot{y}_3 + 6y_3 = u$$

Let us define

$$\mathbf{z} = \begin{bmatrix}z_1\\ z_2\\ z_3\end{bmatrix} = \begin{bmatrix}y_3\\ \dot{y}_3\\ \ddot{y}_3\end{bmatrix}$$

Then, we obtain the desired result:

$$\begin{bmatrix}\dot{z}_1\\ \dot{z}_2\\ \dot{z}_3\end{bmatrix} = \begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ -6 & -11 & -6\end{bmatrix}\begin{bmatrix}z_1\\ z_2\\ z_3\end{bmatrix} + \begin{bmatrix}0\\ 0\\ 1\end{bmatrix}[u]$$

Problem A-4-11 Obtain a state space equation of the following transfer function

$$\frac{X(s)}{U(s)} = \frac{K(s-a)(s-b)}{(s-\lambda_1)(s-\lambda_2)(s-\lambda_3)}$$

where the poles λ_1, λ_2, and λ_3 are distinct.

Solution

$$\frac{X(s)}{U(s)} = \frac{p_{11}}{s-\lambda_1} + \frac{p_{12}}{s-\lambda_2} + \frac{p_{13}}{s-\lambda_3}$$

$$\frac{sX(s)}{U(s)} = K + \frac{p_{21}}{s-\lambda_1} + \frac{p_{22}}{s-\lambda_2} + \frac{p_{23}}{s-\lambda_3}$$

$$\frac{s^2X(s)}{U(s)} = Ks + K(\lambda_1+\lambda_2+\lambda_3-a-b) + \frac{p_{31}}{s-\lambda_1} + \frac{p_{32}}{s-\lambda_2} + \frac{p_{33}}{s-\lambda_3}$$

By defining

$$X_1(s) = X(s)$$
$$X_2(s) = sX(s) - KU(s)$$
$$X_3(s) = s^2X(s) - KsU(s) - K(\lambda_1+\lambda_2+\lambda_3-a-b)U(s)$$

we obtain

$$\begin{bmatrix} \frac{X_1(s)}{U(s)} \\ \frac{X_2(s)}{U(s)} \\ \frac{X_3(s)}{U(s)} \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} \begin{bmatrix} \frac{1}{s-\lambda_1} \\ \frac{1}{s-\lambda_2} \\ \frac{1}{s-\lambda_3} \end{bmatrix}$$

The simplest state space representation of the system is

$$\begin{bmatrix} \frac{Y_1(s)}{U(s)} \\ \frac{Y_2(s)}{U(s)} \\ \frac{Y_3(s)}{U(s)} \end{bmatrix} = \begin{bmatrix} \frac{1}{s-\lambda_1} \\ \frac{1}{s-\lambda_2} \\ \frac{1}{s-\lambda_3} \end{bmatrix}$$

or

$$\dot{y}_1 = \lambda_1 y_1 + u$$
$$\dot{y}_2 = \lambda_2 y_2 + u$$
$$\dot{y}_3 = \lambda_3 y_3 + u$$

A transformation matrix **P** is given by

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

and x is obtained as

$$x = x_1 = p_{11}y_1 + p_{12}y_2 + p_{13}y_3$$

The initial condition terms are given by

$$\begin{bmatrix} y_1(0) \\ y_2(0) \\ y_3(0) \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}^{-1} \begin{bmatrix} x_1(0) \\ x_2(0) \\ x_3(0) \end{bmatrix}$$

where

$$\begin{bmatrix} x_1(0) \\ x_2(0) \\ x_3(0) \end{bmatrix} = \begin{bmatrix} x(0) \\ \dot{x}(0) - Ku(0) \\ \ddot{x}(0) - K\dot{u}(0) - K(\lambda_1 + \lambda_2 + \lambda_3 - a - b)u(0) \end{bmatrix}$$

(Notice that, although the transfer function assumes the zero initial conditions, once the state space equation is obtained we can relax the zero initial conditions to arbitrary initial conditions.)

Problem A-4-12 Obtain a state space equation for the following transfer function:

$$\frac{X(s)}{U(s)} = \frac{s+2}{(s+1)^2(s+3)}$$

Solution In the partial-fraction form, the given transfer function becomes

$$\frac{X(s)}{U(s)} = \frac{1/2}{(s+1)^2} + \frac{1/4}{s+1} + \frac{-1/4}{s+3}$$

Multiply $X(s)/U(s)$ by s and expand it into the partial-fraction form

$$\frac{sX(s)}{U(s)} = \frac{s(s+2)}{(s+1)^2(s+3)} = \frac{-1/2}{(s+1)^2} + \frac{1/4}{s+1} + \frac{3/4}{s+3}$$

Multiply $X(s)/U(s)$ by s^2 and divide the resulting fraction

$$\frac{s^2X(s)}{U(s)} = \frac{s^2(s+2)}{(s+1)^2(s+3)} = 1 - \frac{3s^2+7s+3}{(s+1)^2(s+3)}$$

Then rearrange

$$\frac{s^2X(s) - U(s)}{U(s)} = \frac{-3s^2-7s-3}{(s+1)^2(s+3)}$$
$$= \frac{1/2}{(s+1)^2} + \frac{-3/4}{s+1} + \frac{-9/4}{s+3}$$

By defining

$$X_1(s) = X(s)$$
$$X_2(s) = sX(s)$$
$$X_3(s) = s^2X(s) - U(s)$$

we obtain

$$\begin{bmatrix} \frac{X_1(s)}{U(s)} \\ \frac{X_2(s)}{U(s)} \\ \frac{X_3(s)}{U(s)} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{2} & \frac{1}{4} & \frac{3}{4} \\ \frac{1}{2} & -\frac{3}{4} & -\frac{9}{4} \end{bmatrix} \begin{bmatrix} \frac{1}{(s+1)^2} \\ \frac{1}{s+1} \\ \frac{1}{s+3} \end{bmatrix}$$

By the following transformation from **x** to **y**

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{2} & \frac{1}{4} & \frac{3}{4} \\ \frac{1}{2} & -\frac{3}{4} & -\frac{9}{4} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

we obtain

$$\frac{Y_1(s)}{U(s)} = \frac{1}{(s+1)^2}$$
$$\frac{Y_2(s)}{U(s)} = \frac{1}{s+1}$$
$$\frac{Y_3(s)}{U(s)} = \frac{1}{s+3}$$

or in the state space representation

$$\dot{y}_1 = -y_1 + y_2$$
$$\dot{y}_2 = -y_2 + u$$
$$\dot{y}_3 = -3y_3 + u$$

The initial condition terms are

$$\begin{bmatrix} y_1(0) \\ y_2(0) \\ y_3(0) \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{2} & \frac{1}{4} & \frac{3}{4} \\ \frac{1}{2} & -\frac{3}{4} & -\frac{9}{4} \end{bmatrix}^{-1} \begin{bmatrix} x_1(0) \\ x_2(0) \\ x_3(0) \end{bmatrix}$$

where

$$x_1(0) = x(0)$$
$$x_2(0) = \dot{x}(0)$$
$$x_3(0) = \ddot{x}(0) - u(0)$$

Problem A-4-13 Obtain a state space representation of the following difference equation:

$$x(k+3) + 5x(k+2) + 7x(k+1) + 3x(k) = 0$$

Solution Let us put

$$x_1(k) = x(k)$$
$$x_2(k) = x_1(k+1)$$
$$x_3(k) = x_2(k+1)$$

then, we obtain

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k)$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -3 & -7 & -5 \end{bmatrix}$$

Problem A-4-14 Rewrite the following scalar difference equation

$$x(k+2) + 3x(k+1) + 2x(k) = 5u(k+1) + 3u(k)$$

in the form of a state space equation.

Solution The equation

$$x(k+2) + a_1(k)x(k+1) + a_2(k)x(k) = b_1(k)u(k+1) + b_2(k)u(k)$$

can be written as

$$x_1(k+1) = x_2(k) + b_1(k-1)u(k)$$
$$x_2(k+1) = -a_2(k)x_1(k) - a_1(k)x_2(k) + [b_2(k) - a_1(k)b_1(k-1)]\,u(k)$$

Thus,

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -a_2(k) & -a_1(k) \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} + \begin{bmatrix} b_1(k-1) \\ b_2(k) - a_1(k)b_1(k-1) \end{bmatrix} [u(k)]$$

In the present equation,

$$a_1(k) = 3, \qquad a_2(k) = 2$$
$$b_1(k) = b_1(k-1) = 5, \qquad b_2(k) = 3$$

Thus, we obtain

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} + \begin{bmatrix} 5 \\ -12 \end{bmatrix} [u(k)]$$

Problem A-4-15 Obtain a state space equation for the following difference equation:

$$x(k+3) + 5x(k+2) + 7x(k+1) + 3x(k) = u(k+1) + 2u(k)$$

Solution Let us define

$$x_1(k) = x(k) - c_0 u(k)$$
$$x_1(k+1) = x_2(k) + c_1 u(k)$$
$$x_2(k+1) = x_3(k) + c_2 u(k)$$
$$x_3(k+1) = -3x_1(k) - 7x_2(k) - 5x_3(k) + c_3 u(k)$$

Then we obtain

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{C}u(k) \tag{4-110}$$

where

$$\mathbf{x}(k+1) = \begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ x_3(k+1) \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -3 & -7 & -5 \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} c_1(k) \\ c_2(k) \\ c_3(k) \end{bmatrix}$$

From Eq. (4–102) we find

$$c_0 = 0$$
$$c_1 = 0$$
$$c_2 = 1$$
$$c_3 = -3$$

The roots of the characteristic equation $|\mathbf{A} - \lambda\mathbf{I}| = 0$ are found to be $-1, -1, -3$. In order to simplify Eq. (4–110) let

$$\mathbf{x}(k) = \mathbf{P}\mathbf{y}(k)$$

where

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 1 \\ \lambda_1 & 1 & \lambda_3 \\ \lambda_1^2 & 2\lambda_1 & \lambda_3^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & -3 \\ 1 & -2 & 9 \end{bmatrix}$$

Then Eq. (4–110) becomes

$$\mathbf{y}(k+1) = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}\mathbf{y}(k) + \mathbf{P}^{-1}\mathbf{C}u(k)$$

where

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -3 \end{bmatrix}$$

and

$$\mathbf{P}^{-1}\mathbf{C} = \begin{bmatrix} \frac{3}{4} & -\frac{1}{2} & -\frac{1}{4} \\ \frac{3}{2} & 2 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ -3 \end{bmatrix} = \begin{bmatrix} \frac{1}{4} \\ \frac{1}{2} \\ -\frac{1}{4} \end{bmatrix}$$

By the use of the following transformation

$$\mathbf{y}(k) = \mathbf{V}\mathbf{z}(k)$$

where

$$\mathbf{V} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & -\frac{1}{4} \end{bmatrix}$$

we obtain

$$\mathbf{V}^{-1}\mathbf{P}^{-1}\mathbf{A}\mathbf{P}\mathbf{V} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -3 \end{bmatrix}$$

and

$$\mathbf{V}^{-1}\mathbf{P}^{-1}\mathbf{C} = \begin{bmatrix} 2 & -1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -4 \end{bmatrix} \begin{bmatrix} \frac{1}{4} \\ \frac{1}{2} \\ -\frac{1}{4} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

The simplified state space equation is given by

$$\begin{bmatrix} z_1(k+1) \\ z_2(k+1) \\ z_3(k+1) \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} z_1(k) \\ z_2(k) \\ z_3(k) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} [u(k)]$$

where the initial data $\mathbf{z}(0)$ are given by

$$\mathbf{z}(0) = \mathbf{V}^{-1}\mathbf{y}(0) = \mathbf{V}^{-1}\mathbf{P}^{-1}\mathbf{x}(0)$$

and

$$\mathbf{x}(0) = \begin{bmatrix} x_1(0) \\ x_2(0) \\ x_3(0) \end{bmatrix} = \begin{bmatrix} x(0) \\ x(1) \\ x(2) - u(0) \end{bmatrix}$$

In order to obtain $\mathbf{z}(0)$, we must know $x(0)$, $x(1)$, $x(2)$, and $u(0)$.

Notice that in this problem

$$\mathbf{x} = \mathbf{P}\mathbf{V}\mathbf{z}$$

where

$$\mathbf{P}\mathbf{V} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{2} & \frac{1}{4} & \frac{3}{4} \\ \frac{1}{2} & -\frac{3}{4} & -\frac{9}{4} \end{bmatrix}$$

will directly transform Eq. (4–110) into the simplified form. The transformation matrix $\mathbf{PV}$ is the same as the transformation matrix obtained in Prob. A-4-12 by the partial-fraction expansion technique for a similar system

$$\frac{X(s)}{U(s)} = \frac{s+2}{(s+1)^2(s+3)}$$

or

$$(s^3 + 5s^2 + 7s + 3)X(s) = (s+2)U(s)$$

Notice also that in Prob. A-4-12, $x_1(t)$, $x_2(t)$, and $x_3(t)$ are defined by

$$x_1(t) = x(t)$$
$$x_2(t) = \dot{x}(t)$$
$$x_3(t) = \ddot{x}(t) - u(t)$$

whereas in this problem, $x_1(k)$, $x_2(k)$, and $x_3(k)$ are defined by

$$x_1(k) = x(k)$$
$$x_2(k) = x_1(k+1) = x(k+1)$$
$$x_3(k) = x_2(k+1) - c_2u(k) = x(k+2) - u(k)$$

PROBLEMS

Problem B-4-1 Obtain state space equations of the following systems:

(1) $\dddot{x} + 5\ddot{x} + 7\dot{x} + 3x = \dot{u} + 2u$

(2) $\dddot{x} + 5\ddot{x} + 7\dot{x} + 3x = \ddot{u} + 3\dot{u} + 2u$

Problem B-4-2 Given the following state space equation

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ -2 & -3 & 0 \\ -1 & 1 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} [u]$$

Obtain the corresponding scalar differential equation of the form:

$$\dddot{y}_1 + a_1\ddot{y}_1 + a_2\dot{y}_1 + a_3y_1 = b_1\ddot{u} + b_2\dot{u} + b_3u$$

Examine if it is possible to transform the original equation into the form

$$\dddot{y}_3 + a_1\ddot{y}_3 + a_2\dot{y}_3 + a_3y_3 = u$$

Problem B-4-3 Suppose that $\mathbf{A}$ is an $n \times n$ nonsingular matrix and $\mathbf{B}$ is an $n \times 1$ matrix. Suppose also that the column vectors $\mathbf{B}$, $\mathbf{AB}$, . . . , $\mathbf{A}^{n-1}\mathbf{B}$ are linearly independent. Define an $n \times n$ nonsingular matrix $\mathbf{P}$ such that

$$\mathbf{P} = [\mathbf{A}^{n-1}\mathbf{B} \mid \mathbf{A}^{n-2}\mathbf{B} \mid \cdots \mid \mathbf{AB} \mid \mathbf{B}]$$

Show that

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{bmatrix} -a_1 & 1 & 0 & \cdots & 0 & 0 \\ -a_2 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ -a_{n-1} & 0 & 0 & \cdots & 0 & 1 \\ -a_n & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

where $a_1, a_2, \ldots, a_n$ are constants determined by **A** and **B**.

Problem B-4-4 Obtain a state space equation for the following transfer function:

$$\frac{X(s)}{U(s)} = \frac{(s+2)}{(s^2+s+1)(s+1)}$$

Problem B-4-5 Transform the following differential equations into state space equations by use of the partial-fraction expansion technique.

(1) $\dddot{x} + 3\ddot{x} + 3\dot{x} + x = u$

(2) $\dddot{x} + 3\ddot{x} + 3\dot{x} + x = \dot{u} + u$

Problem B-4-6 Derive a state space equation of the system shown in Fig. 4–6. By what transformation can we transform the state space equation into

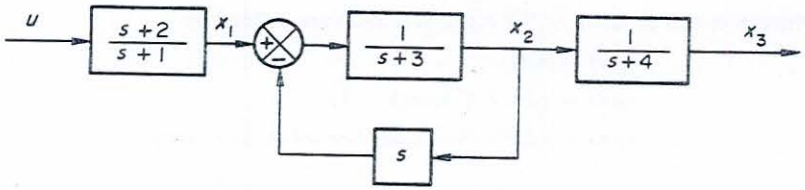


Figure 4-6 Control system.

the simplified form where the coefficient matrix of the state vector is in the diagonal form and the elements of the coefficient matrix of u are unity?

Problem B-4-7 Transform the following difference equation:

$$x(k+2) + 3x(k+1) + 2x(k) = 2u(k+1) + 3u(k)$$

into state space equations where the coefficient matrix **B** of the forcing function u is given by

$$\text{(1)} \quad \mathbf{B} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad \text{(2)} \quad \mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

REFERENCES

4–1 Wonham, W. M., and C. D. Johnson, "Optimal Bang-Bang Control with Quadratic Performance Index," *ASME J. Basic Engineering*, ser. D, **86** (1964), 107–115.

4–2 Ogata, K., Class Notes on Control Theory, Lecture, Yokohama National University, 1960–61.

4–3 Ichikawa, K., "Coordinate Transformation," *Tech. Reports Inst. Automatic Control* (Faculty of Engineering, Nagoya University), **10** (1962), 61–77.

5

Matrix Analysis

5-1 INTRODUCTION

In Chapter 2 we presented preliminary materials on matrix analysis. In this chapter, we present additional materials on matrix analysis which are needed in the rest of this book. Specifically, we shall discuss functions of a square matrix, quadratic forms, and Hermitian forms.

In the succeeding sections of this chapter, we proceed as follows: Section 5–2 presents matrix polynomials, Frobenius theorem, Cayley-Hamilton theorem, the minimal polynomial, and the exponential function of a matrix. Section 5–3 gives Sylvester's interpolation formula. In Section 5–4 we discuss real quadratic forms and Hermitian forms and shall present theorems on real symmetric matrices and Hermitian matrices. We shall also discuss definiteness, semidefiniteness, and indefiniteness of real quadratic forms and Hermitian forms. Finally, Section 5–5 presents Sylvester's criteria for definiteness and semidefiniteness of real quadratic forms and Hermitian forms.

5-2 FUNCTIONS OF SQUARE MATRIX

In this section we shall discuss polynomials in a square matrix, Frobenius theorem, Cayley-Hamilton theorem, the minimal polynomial, and the exponential function of a matrix.

Polynomials in square matrix Let $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_p$ and $\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_q$ be $n \times n$ matrices. Define $f(x)$, $g(x)$, and $h(x)$, where x is a scalar, as follows:

$$f(x) = \sum_{i=0}^{p} \mathbf{A}_i x^i$$

$$g(x) = \sum_{j=0}^{q} \mathbf{B}_j x^j$$

$$h(x) = f(x)g(x) = \sum_{i=0}^{p} \sum_{j=0}^{q} \mathbf{A}_i \mathbf{B}_j x^{i+j}$$

We must be careful in substituting an $n \times n$ matrix $\mathbf{C}$ for x in the preceding equation for $h(x)$, since the multiplication of matrices is not commutative in general. We have the following relationships:

$$\sum_{i=0}^{p} \sum_{j=0}^{q} \mathbf{A}_i \mathbf{B}_j \mathbf{C}^{i+j}$$

$$= \mathbf{A}_0\left(\sum_{j=0}^{q} \mathbf{B}_j \mathbf{C}^j\right) + \mathbf{A}_1\left(\sum_{j=0}^{q} \mathbf{B}_j \mathbf{C}^j\right)\mathbf{C} + \cdots + \mathbf{A}_p\left(\sum_{j=0}^{q} \mathbf{B}_j \mathbf{C}^j\right)\mathbf{C}^p$$

$$= \left(\sum_{i=0}^{p} \mathbf{A}_i \mathbf{C}^i\right)\left(\sum_{j=0}^{q} \mathbf{B}_j \mathbf{C}^j\right) \qquad \text{if } \mathbf{B} \text{ and } \mathbf{C} \text{ commute}$$

$$\sum_{i=0}^{p} \sum_{j=0}^{q} \mathbf{A}_i \mathbf{B}_j \mathbf{C}^{i+j} \neq \left(\sum_{i=0}^{p} \mathbf{A}_i \mathbf{C}^i\right)\left(\sum_{j=0}^{q} \mathbf{B}_j \mathbf{C}^j\right) \qquad \text{if } \mathbf{B} \text{ and } \mathbf{C} \text{ do not commute}$$

Notice, however, that

$$\sum_{i=0}^{p} \sum_{j=0}^{q} \mathbf{A}_i \mathbf{B}_j \mathbf{C}^{i+j} = \mathbf{0} \qquad \text{if} \qquad \sum_{j=0}^{q} \mathbf{B}_j \mathbf{C}^j = \mathbf{0}$$

Also

$$\sum_{i=0}^{p} \sum_{j=0}^{q} \mathbf{C}^{i+j} \mathbf{A}_i \mathbf{B}_j$$

$$= \left(\sum_{i=0}^{p} \mathbf{C}^i \mathbf{A}_i\right)\mathbf{B}_0 + \mathbf{C}\left(\sum_{i=0}^{p} \mathbf{C}^i \mathbf{A}_i\right)\mathbf{B}_1 + \cdots + \mathbf{C}^q\left(\sum_{i=0}^{p} \mathbf{C}^i \mathbf{A}_i\right)\mathbf{B}_q$$

$$= \left(\sum_{i=0}^{p} \mathbf{C}^i \mathbf{A}_i\right)\left(\sum_{j=0}^{q} \mathbf{C}^j \mathbf{B}_j\right) \qquad \text{if } \mathbf{A} \text{ and } \mathbf{C} \text{ commute}$$

$$\sum_{i=0}^{p} \sum_{j=0}^{q} \mathbf{C}^{i+j} \mathbf{A}_i \mathbf{B}_j \neq \left(\sum_{i=0}^{p} \mathbf{C}^i \mathbf{A}_i\right)\left(\sum_{j=0}^{q} \mathbf{C}^j \mathbf{B}_j\right) \qquad \text{if } \mathbf{A} \text{ and } \mathbf{C} \text{ do not commute}$$

Notice that

$$\sum_{i=0}^{p} \sum_{j=0}^{q} \mathbf{C}^{i+j} \mathbf{A}_i \mathbf{B}_j = \mathbf{0} \qquad \text{if} \qquad \sum_{i=0}^{p} \mathbf{C}^i \mathbf{A}_i = \mathbf{0}$$

We shall use these results occasionally.

Frobenius theorem We shall next present Frobenius theorem concerning the eigenvalues of a polynomial in an $n \times n$ matrix $\mathbf{A}$.

Theorem 5-1 (*Frobenius theorem*) Suppose that the eigenvalues of an $n \times n$ matrix $\mathbf{A}$ are $\lambda_1, \lambda_2, \ldots, \lambda_n$. Let $p(x)$ be an arbitrary polynomial in x, or

$$p(x) = a_0 x^p + a_1 x^{p-1} + \cdots + a_p$$

Then the eigenvalues of the matrix polynomial $p(\mathbf{A})$ where

$$p(\mathbf{A}) = a_0 \mathbf{A}^p + a_1 \mathbf{A}^{p-1} + \cdots + a_p \mathbf{I}$$

are $p(\lambda_1), p(\lambda_2), \ldots, p(\lambda_n)$.

Proof Let the roots of $p(x) - \lambda = 0$ be $\mu_1, \mu_2, \ldots, \mu_p$. Then

$$p(x) - \lambda = a_0(x - \mu_1)(x - \mu_2) \cdots (x - \mu_p) = a_0 \prod_{i=1}^{p} (x - \mu_i)$$

Noting that

$$a_0(\mu_1 + \mu_2 + \cdots + \mu_p) = -a_1$$

$$a_0(\mu_1\mu_2 + \mu_1\mu_3 + \cdots + \mu_{p-1}\mu_p) = a_2$$

$$\cdots \qquad \cdots \qquad \cdots$$

$$a_0\mu_1\mu_2 \cdots \mu_p = (-)^p(a_p - \lambda)$$

we have

$$p(\mathbf{A}) - \lambda\mathbf{I} = a_0\mathbf{A}^p + a_1\mathbf{A}^{p-1} + \cdots + (a_p - \lambda)\mathbf{I}$$

$$= a_0(\mathbf{A} - \mu_1\mathbf{I})(\mathbf{A} - \mu_2\mathbf{I}) \cdots (\mathbf{A} - \mu_p\mathbf{I})$$

Taking the determinant of $p(\mathbf{A}) - \lambda\mathbf{I}$

$$|p(\mathbf{A}) - \lambda\mathbf{I}| = a_0^n |\mathbf{A} - \mu_1\mathbf{I}||\mathbf{A} - \mu_2\mathbf{I}| \cdots |\mathbf{A} - \mu_p\mathbf{I}|$$

Since

$$|\mathbf{A} - \lambda\mathbf{I}| = (\lambda_1 - \lambda)(\lambda_2 - \lambda) \cdots (\lambda_n - \lambda)$$

we obtain

$$|p(\mathbf{A}) - \lambda\mathbf{I}| = a_0^n\left[\prod_{j=1}^{n} (\lambda_j - \mu_1)\right]\left[\prod_{j=1}^{n} (\lambda_j - \mu_2)\right] \cdots \left[\prod_{j=1}^{n} (\lambda_j - \mu_p)\right]$$

$$= \prod_{i=1}^{p} a_0^n \prod_{j=1}^{n} (\lambda_j - \mu_i)$$

$$= \prod_{j=1}^{n}\left[a_0 \prod_{i=1}^{p} (\lambda_j - \mu_i)\right]$$

$$= \prod_{j=1}^{n} [p(\lambda_j) - \lambda]$$

$$= [p(\lambda_1) - \lambda][p(\lambda_2) - \lambda] \cdots [p(\lambda_n) - \lambda]$$

This proves that the eigenvalues of $p(\mathbf{A})$ are $p(\lambda_1), p(\lambda_2), \ldots, p(\lambda_n)$.

Cayley-Hamilton theorem The Cayley-Hamilton theorem plays an important role in simplifying polynomials in an $n \times n$ matrix $\mathbf{A}$. It states that $\mathbf{A}$ satisfies its characteristic equation.

We shall first give an illustrative example and then present the theorem. Consider the following matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 0 \\ 3 & -1 & -2 \\ 1 & 0 & -3 \end{bmatrix}$$

The characteristic equation is

$$|\mathbf{A} - \lambda\mathbf{I}| = -\lambda^3 - 3\lambda^2 + 7\lambda + 17 = 0$$

Notice that $(\mathbf{A} - \lambda\,\mathbf{I})$ and adj $(\mathbf{A} - \lambda\,\mathbf{I})$ are polynomials in λ of degree one and degree two, respectively. Namely,

$$\mathbf{A} - \lambda\mathbf{I} = -\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\lambda + \begin{bmatrix} 1 & 2 & 0 \\ 3 & -1 & -2 \\ 1 & 0 & -3 \end{bmatrix}$$

and

$$\text{adj}\,(\mathbf{A} - \lambda\mathbf{I}) = \begin{bmatrix} \lambda^2 + 4\lambda + 3 & 2\lambda + 6 & -4 \\ 3\lambda + 7 & \lambda^2 + 2\lambda - 3 & -2\lambda + 2 \\ \lambda + 1 & 2 & \lambda^2 - 7 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\lambda^2 + \begin{bmatrix} 4 & 2 & 0 \\ 3 & 2 & -2 \\ 1 & 0 & 0 \end{bmatrix}\lambda + \begin{bmatrix} 3 & 6 & -4 \\ 7 & -3 & 2 \\ 1 & 2 & -7 \end{bmatrix}$$

Observe that adj $(\mathbf{A} - \lambda\mathbf{I})$ is a polynomial in λ of degree $n - 1$, where $n = 3$ in this example. If we multiply $(\mathbf{A} - \lambda\mathbf{I})$ by adj $(\mathbf{A} - \lambda\mathbf{I})$ we obtain

$$(\mathbf{A} - \lambda\mathbf{I})\ \text{adj}\ (\mathbf{A} - \lambda\mathbf{I})$$

$$= -\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\lambda^3 + \begin{bmatrix} -3 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & -3 \end{bmatrix}\lambda^2$$

$$+\begin{bmatrix} 7 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 7 \end{bmatrix}\lambda + \begin{bmatrix} 17 & 0 & 0 \\ 0 & 17 & 0 \\ 0 & 0 & 17 \end{bmatrix}$$

$$= -\lambda^3\mathbf{I} - 3\lambda^2\mathbf{I} + 7\lambda\mathbf{I} + 17\mathbf{I}$$

$$= (-\lambda^3 - 3\lambda^2 + 7\lambda + 17)\,\mathbf{I} \qquad (5\text{–}1)$$

Observe that if we substitute $\mathbf{A}$ for λ in Eq. (5–1), then

$$-\mathbf{A}^3 - 3\mathbf{A}^2 + 7\mathbf{A} + 17\mathbf{I}$$

$$= -\begin{bmatrix} 3 & 14 & 12 \\ 27 & -11 & -38 \\ 13 & -6 & -31 \end{bmatrix} - 3\begin{bmatrix} 7 & 0 & -4 \\ -2 & 7 & 8 \\ -2 & 2 & 9 \end{bmatrix}$$

$$+7\begin{bmatrix} 1 & 2 & 0 \\ 3 & -1 & -2 \\ 1 & 0 & -3 \end{bmatrix} + 17\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Thus, if $\mathbf{A}$ is substituted for λ in the characteristic polynomial multiplied by the identity matrix, then the result is zero. This is an illustrative example of the following theorem:

Theorem 5-2 (*Cayley-Hamilton theorem*) Let $\mathbf{A}$ be an $n \times n$ matrix and its characteristic equation be

$$|\mathbf{A} - \lambda\mathbf{I}| = (-)^n(\lambda^n + c_1\lambda^{n-1} + \cdots + c_{n-2}\lambda^2 + c_{n-1}\lambda + c_n) = 0$$

The matrix $\mathbf{A}$ satisfies its characteristic equation, or

$$\mathbf{A}^n + c_1\mathbf{A}^{n-1} + \cdots + c_{n-2}\mathbf{A}^2 + c_{n-1}\mathbf{A} + c_n\mathbf{I} = \mathbf{0}$$

Proof To prove the theorem we first note that adj $(\mathbf{A} - \lambda\mathbf{I})$ is a polynomial in λ of degree $n - 1$, namely,

$$\text{adj}\,(\mathbf{A} - \lambda\mathbf{I}) = \mathbf{B}_1\lambda^{n-1} + \mathbf{B}_2\lambda^{n-2} + \cdots + \mathbf{B}_{n-2}\lambda^2 + \mathbf{B}_{n-1}\lambda + \mathbf{B}_n$$

where $$\mathbf{B}_1 = (-)^{n-1}\,\mathbf{I}$$

Note also that

$$(\mathbf{A} - \lambda\mathbf{I})\,\text{adj}\,(\mathbf{A} - \lambda\mathbf{I}) = [\text{adj}\,(\mathbf{A} - \lambda\mathbf{I})](\mathbf{A} - \lambda\mathbf{I}) = |\mathbf{A} - \lambda\mathbf{I}|\,\mathbf{I}$$

Hence we obtain

$$|\mathbf{A} - \lambda\mathbf{I}|\,\mathbf{I} = (-)^n(\mathbf{I}\lambda^n + c_1\mathbf{I}\lambda^{n-1} + \cdots + c_{n-2}\mathbf{I}\lambda^2 + c_{n-1}\mathbf{I}\lambda + c_n\mathbf{I})$$

$$= (-\mathbf{I}\lambda + \mathbf{A})(\mathbf{B}_1\lambda^{n-1} + \mathbf{B}_2\lambda^{n-2} + \cdots + \mathbf{B}_{n-2}\lambda^2 + \mathbf{B}_{n-1}\lambda + \mathbf{B}_n)$$

$$= (\mathbf{B}_1\lambda^{n-1} + \mathbf{B}_2\lambda^{n-2} + \cdots + \mathbf{B}_{n-2}\lambda^2 + \mathbf{B}_{n-1}\lambda + \mathbf{B}_n)(-\mathbf{I}\lambda + \mathbf{A})$$

From this equation, we see that $\mathbf{A}$ and $\mathbf{B}_i$ $(i = 0, 1, 2, \ldots, n - 1)$ commute. Hence the product of $(\mathbf{A} - \lambda\mathbf{I})$ and adj $(\mathbf{A} - \lambda\mathbf{I})$ becomes zero if either one of these is zero. If $\mathbf{A}$ is substituted for λ in this last equation, then clearly $\mathbf{A} - \lambda\mathbf{I}$ becomes zero. Hence,

$$\mathbf{A}^n + c_1\mathbf{A}^{n-1} + \cdots + c_{n-2}\mathbf{A}^2 + c_{n-1}\mathbf{A} + c_n\mathbf{I} = \mathbf{0}$$

This proves the theorem.

Minimal polynomial As stated in Theorem 5–2, every $n \times n$ matrix $\mathbf{A}$ satisfies its own characteristic equation. The characteristic equation is not, however, necessarily the scalar equation of least degree that $\mathbf{A}$ satisfies.

The least degree polynomial having $\mathbf{A}$ as a root is called the minimal polynomial; namely, the minimal polynomial of an $n \times n$ matrix $\mathbf{A}$ is defined as the polynomial $\phi(\lambda)$ of least degree

$$\phi(\lambda) = \lambda^m + a_1\lambda^{m-1} + \cdots + a_m, \qquad m \leqslant n$$

such that $\phi(\mathbf{A}) = \mathbf{0}$, or

$$\phi(\mathbf{A}) = \mathbf{A}^m + a_1\mathbf{A}^{m-1} + \cdots + a_m\mathbf{I} = \mathbf{0}$$

The minimal polynomial plays an important role in the computation of polynomials in an $n \times n$ matrix.

Theorem 5-3 Let $d(\lambda)$, a polynomial in λ, be the greatest common divisor of all the $(n-1)$-rowed minors of an $n \times n$ matrix $(\mathbf{A} - \lambda\mathbf{I})$. If the coefficient of the highest degree term in λ of $d(\lambda)$ is chosen as $(-1)^n$, then the minimal polynomial $\phi(\lambda)$ is given by

$$\phi(\lambda) = \frac{|\mathbf{A} - \lambda\mathbf{I}|}{d(\lambda)}$$

Proof By assumption, the greatest common divisor of the matrix adj $(\mathbf{A} - \lambda\mathbf{I})$ is $d(\lambda)$. Therefore

$$\text{adj}\,(\mathbf{A} - \lambda\mathbf{I}) = d(\lambda)\mathbf{B}(\lambda)$$

where the greatest common divisor of the n^2 elements (which are functions of λ) of $\mathbf{B}(\lambda)$ is unity. Since

$$(\mathbf{A} - \lambda\mathbf{I})\,\text{adj}\,(\mathbf{A} - \lambda\mathbf{I}) = |\mathbf{A} - \lambda\mathbf{I}|\mathbf{I}$$

we obtain

$$d(\lambda)(\mathbf{A} - \lambda\mathbf{I})\mathbf{B}(\lambda) = |\mathbf{A} - \lambda\mathbf{I}|\mathbf{I}$$

from which we find that $|\mathbf{A} - \lambda\mathbf{I}|$ is divisible by $d(\lambda)$. Let us put

$$|\mathbf{A} - \lambda\mathbf{I}| = d(\lambda)\psi(\lambda)$$

Then the coefficient of the highest degree term in λ of $\psi(\lambda)$ is unity. Since we have

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{B}(\lambda) = \psi(\lambda)\mathbf{I}$$

it follows that

$$\psi(\mathbf{A}) = \mathbf{0}$$

Note that $\psi(\lambda)$ can be written as

$$\psi(\lambda) = g(\lambda)\phi(\lambda) + \alpha(\lambda)$$

where $\alpha(\lambda)$ is of lower degree than $\phi(\lambda)$. If $\psi(\mathbf{A}) = \mathbf{0}$, then $\alpha(\mathbf{A}) = \mathbf{0}$. Since $\phi(\lambda)$ is the minimal polynomial, $\alpha(\lambda)$ must be equal to zero, or

$$\psi(\lambda) = g(\lambda)\phi(\lambda)$$

Note that because $\phi(\mathbf{A}) = \mathbf{0}$, we can write

$$\phi(\lambda)\mathbf{I} = (\mathbf{A} - \lambda\mathbf{I})\mathbf{C}(\lambda)$$

Hence

$$\psi(\lambda)\mathbf{I} = g(\lambda)\phi(\lambda)\mathbf{I} = g(\lambda)(\mathbf{A} - \lambda\mathbf{I})\mathbf{C}(\lambda)$$

and we obtain

$$\mathbf{B}(\lambda) = g(\lambda)\mathbf{C}(\lambda)$$

Note that the greatest common divisor of n^2 elements of $\mathbf{B}(\lambda)$ is unity. Hence,

$$g(\lambda) = 1$$

Therefore,

$$\psi(\lambda) = \phi(\lambda)$$

it follows that

$$\phi(\lambda) = \frac{|\mathbf{A} - \lambda\mathbf{I}|}{d(\lambda)}$$

This proves the theorem.

It can be shown that similar matrices have the same minimal polynomial. Namely, for a nonsingular matrix $\mathbf{T}$, the minimal polynomial of $\mathbf{T}^{-1}\mathbf{A}\mathbf{T}$ and that of $\mathbf{A}$ are the same. (For the proof, see Prob. A-5-3.)

The minimal polynomial $\phi(\lambda)$ of an $n \times n$ matrix $\mathbf{A}$ can be determined by the following procedure:

1. Form adj $(\mathbf{A} - \lambda\mathbf{I})$ and write the elements of adj $(\mathbf{A} - \lambda\mathbf{I})$ as factored polynomials in λ.
2. Determine $d(\lambda)$ as the greatest common divisor of all the elements of adj $(\mathbf{A} - \lambda\mathbf{I})$. Choose the coefficient of the highest degree term in λ of $d(\lambda)$ be $(-1)^n$. If there is no common divisor, $d(\lambda) = (-1)^n$.
3. The minimal polynomial $\phi(\lambda)$ is then given as $|\mathbf{A} - \lambda\mathbf{I}|$ divided by $d(\lambda)$.

Whenever $\mathbf{A}$ has distinct eigenvalues, apart from the constant factor $(-)^n$ the minimal polynomial is identical with the characteristic polynomial.

Note that if the multiple eigenvalues of $\mathbf{A}$ are linked in a Jordan chain, the minimal polynomial and the characteristic polynomial are identical. As an example, consider

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 4 \\ 0 & 2 & 0 \\ 0 & 3 & 1 \end{bmatrix}$$

The characteristic polynomial is given by

$$|\mathbf{A} - \lambda\mathbf{I}| = \begin{vmatrix} 2-\lambda & 1 & 4 \\ 0 & 2-\lambda & 0 \\ 0 & 3 & 1-\lambda \end{vmatrix} = (2-\lambda)^2(1-\lambda)$$

Thus the eigenvalues of $\mathbf{A}$ are 2, 2, and 1. It can be shown that the Jordan canonical form of $\mathbf{A}$ is

$$\begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and the multiple eigenvalues are linked in the Jordan chain as shown. In the following we shall illustrate that the minimal polynomial and the characteristic polynomial of this matrix $\mathbf{A}$ are the same. To compute the minimal polynomial let us obtain adj $(\mathbf{A} - \lambda\mathbf{I})$. It is given by

$$\text{adj}\,(\mathbf{A} - \lambda\mathbf{I}) = \begin{bmatrix} (2-\lambda)(1-\lambda) & (\lambda + 11) & -4(2-\lambda) \\ 0 & (2-\lambda)(1-\lambda) & 0 \\ 0 & -3(2-\lambda) & (2-\lambda)^2 \end{bmatrix}$$

Notice that there is no common divisor of all the elements of adj $(\mathbf{A} - \lambda\mathbf{I})$. Hence $d(\lambda) = (-1)^3 = -1$. Apart from the minus sign $\phi(\lambda)$ is identical with the characteristic polynomial, or

$$\begin{aligned} \phi(\lambda) &= -|\mathbf{A} - \lambda\mathbf{I}| = -(2-\lambda)^2(1-\lambda) \\ &= \lambda^3 - 5\lambda^2 + 8\lambda - 4 \end{aligned}$$

A simple calculation proves that

$$\mathbf{A}^3 - 5\mathbf{A}^2 + 8\mathbf{A} - 4\mathbf{I} = \mathbf{0}$$

but

$$\mathbf{A}^2 - 3\mathbf{A} + 2\mathbf{I} \neq \mathbf{0}$$

Note that if the multiple eigenvalues of $\mathbf{A}$ are not linked in a Jordan chain, the minimal polynomial is of lower degree than the characteristic polynomial. Consider the following example:

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 3 & 1 \end{bmatrix}$$

The characteristic polynomial is given by

$$|\mathbf{A} - \lambda\mathbf{I}| = \begin{bmatrix} 2-\lambda & 0 & 0 \\ 0 & 2-\lambda & 0 \\ 0 & 3 & 1-\lambda \end{bmatrix} = (2-\lambda)^2(1-\lambda)$$

A simple computation reveals that $\mathbf{A}$ has three eigenvectors, and the Jordan canonical form of $\mathbf{A}$ is given by

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Thus, the multiple eigenvalues are not linked. To obtain the minimal polynomial we first compute adj $(\mathbf{A} - \lambda\mathbf{I})$, or

$$\text{adj}\,(\mathbf{A} - \lambda\mathbf{I}) = \begin{bmatrix} (2-\lambda)(1-\lambda) & 0 & 0 \\ 0 & (2-\lambda)(1-\lambda) & 0 \\ 0 & -3(2-\lambda) & (2-\lambda)^2 \end{bmatrix}$$

from which it is evident that

$$d(\lambda) = 2 - \lambda$$

Hence

$$\begin{aligned} \phi(\lambda) &= \frac{|\mathbf{A} - \lambda\mathbf{I}|}{d(\lambda)} = \frac{(2-\lambda)^2(1-\lambda)}{2-\lambda} \\ &= \lambda^2 - 3\lambda + 2 \end{aligned}$$

As a check, let us compute $\phi(\mathbf{A})$.

$$\begin{aligned} \phi(\mathbf{A}) &= \mathbf{A}^2 - 3\mathbf{A} + 2\mathbf{I} \\ &= \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 9 & 1 \end{bmatrix} - 3\begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 3 & 1 \end{bmatrix} + 2\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{aligned}$$

For the given $\mathbf{A}$ the degree of the minimal polynomial is one lower than that of the characteristic polynomial.

Inverse of A Notice that, by the use of the minimal polynomial, the inverse of a nonsingular matrix $\mathbf{A}$ can be expressed as a polynomial in $\mathbf{A}$ with scalar coefficients. Consider a nonsingular matrix $\mathbf{A}$ and its minimal polynomial $\phi(\mathbf{A})$ where

$$\phi(\mathbf{A}) = \mathbf{A}^m + a_1\mathbf{A}^{m-1} + \cdots + a_m\mathbf{I} = \mathbf{0}$$

For a nonsingular matrix, $a_m \neq 0$. Hence,

$$\mathbf{I} = -\frac{1}{a_m}(\mathbf{A}^m + a_1\mathbf{A}^{m-1} + \cdots + a_{m-1}\mathbf{A})$$

Premultiplying with $\mathbf{A}^{-1}$ we obtain

$$\mathbf{A}^{-1} = -\frac{1}{a_m}(\mathbf{A}^{m-1} + a_1\mathbf{A}^{m-2} + \cdots + a_{m-1}\mathbf{I}) \tag{5–2}$$

Thus, $\mathbf{A}^{-1}$ can be numerically computed according to Eq. (5–2).

We shall next present a theorem on infinite series in an $n \times n$ matrix A.

Theorem 5-4 If the power series

$$f(z) = \sum_{k=0}^{\infty} c_k z^k$$

in a complex variable z converges everywhere, then the matrix power series

$$\sum_{k=0}^{\infty} c_k \mathbf{A}^k$$

in an $n \times n$ matrix $\mathbf{A}$ converges absolutely.

Proof We shall first prove that the infinite series $\sum_{k=0}^{\infty} c_k \mathbf{A}^k$ is absolutely convergent if and only if $\sum_{k=0}^{\infty} \| c_k \mathbf{A}^k \|$ is convergent. Define $\mathbf{A}^k = (a_{ij}^k)$. If $\sum_{k=0}^{\infty} c_k \mathbf{A}^k$ is absolutely convergent, then there exists a constant M such that

$$\sum_{k=0}^{N} c_k |a_{ij}^k| \leqslant \sum_{k=0}^{N} |c_k| |a_{ij}^k| < M$$

$$(N \geqslant 0\,;\ i, j = 1, 2, \ldots, n)$$

where M is independent of i, j, and N. Since

$$\sum_{k=0}^{N} \| c_k \mathbf{A}^k \| \leqslant \sum_{k=0}^{N} |c_k| \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}^k| < n^2 M$$

we conclude that $\sum_{k=0}^{\infty} \| c_k \mathbf{A}^k \|$ is convergent.

Conversely, if $\sum_{k=0}^{\infty} \| c_k \mathbf{A}^k \|$ is convergent, then the series consisting of the (i, j)th elements of $\sum_{k=0}^{\infty} c_k \mathbf{A}^k$ is absolutely convergent, since

$$|c_k a_{ij}^k| \leqslant \| c_k \mathbf{A}^k \|$$

Hence we have proved that the series $\sum_{k=0}^{\infty} c_k \mathbf{A}^k$ is absolutely convergent if and only if $\sum_{k=0}^{\infty} \| c_k \mathbf{A}^k \|$ is convergent.

Now by the assumption in the theorem we have

$$\sum_{k=0}^{\infty} \| c_k \mathbf{A}^k \| \leqslant \sum_{k=0}^{\infty} |c_k| (\| \mathbf{A} \|)^k < M'$$

where M' is a constant. It follows that $\sum_{k=0}^{\infty} \| c_k \mathbf{A}^k \|$ is convergent. Hence the theorem.

Exponential function of square matrix In the scalar case e^z is defined by

$$e^z = 1 + z + \frac{1}{2!} z^2 + \frac{1}{3!} z^3 + \cdots$$

Since this power series

$$\sum_{k=0}^{\infty} \frac{1}{k!} z^k$$

converges everywhere, the matrix power series

$$\sum_{k=0}^{\infty} \frac{1}{k!} \mathbf{A}^k$$

converges absolutely for any square matrix $\mathbf{A}$. The exponential function of a matrix can thus be defined for every square matrix $\mathbf{A}$ by

$$e^{\mathbf{A}} = \mathbf{I} + \mathbf{A} + \frac{1}{2!} \mathbf{A}^2 + \frac{1}{3!} \mathbf{A}^3 + \cdots$$

where we used $\mathbf{I}$ for $\mathbf{A}^0$. Similarly, since $\cos z$ and $\sin z$ converge everywhere, we have

$$\cos \mathbf{A} = \mathbf{I} - \frac{1}{2!} \mathbf{A}^2 + \frac{1}{4!} \mathbf{A}^4 - \cdots$$

$$\sin \mathbf{A} = \mathbf{A} - \frac{1}{3!} \mathbf{A}^3 + \frac{1}{5!} \mathbf{A}^5 - \cdots$$

By the definition of the exponential function of $\mathbf{A}$, the exponential function of $\mathbf{A}t$ can be written as

$$e^{\mathbf{A}t} = \mathbf{I} + \mathbf{A}t + \frac{1}{2!} \mathbf{A}^2 t^2 + \cdots + \frac{1}{n!} \mathbf{A}^n t^n + \cdots$$

This series converges absolutely for all finite t. (Refer to Prob. A-5-5.)

Let us next examine the derivative of $e^{\mathbf{A}t}$ with respect to t. Differentiation of $e^{\mathbf{A}t}$ with respect to t yields

$$\begin{aligned}
\frac{d}{dt}(e^{\mathbf{A}t}) &= \mathbf{A} + \mathbf{A}^2 t + \cdots + \mathbf{A}^n \frac{n t^{n-1}}{n!} + \cdots \\
&= \mathbf{A}\left(\mathbf{I} + \mathbf{A}t + \cdots + \mathbf{A}^{n-1} \frac{t^{n-1}}{(n-1)!} + \cdots\right) \\
&= \mathbf{A} e^{\mathbf{A}t} \\
&= \left(\mathbf{I} + \mathbf{A}t + \cdots + \mathbf{A}^{n-1} \frac{t^{n-1}}{(n-1)!} + \cdots\right)\mathbf{A} \\
&= e^{\mathbf{A}t} \mathbf{A}
\end{aligned}$$

This relationship is identical with the scalar case. Let us examine $e^{\mathbf{A}t}\, e^{\mathbf{A}s}$.

$$\begin{aligned}
e^{\mathbf{A}t} e^{\mathbf{A}s} &= \left(\sum_{n=0}^{\infty} \frac{\mathbf{A}^n t^n}{n!}\right)\left(\sum_{n=0}^{\infty} \frac{\mathbf{A}^n s^n}{n!}\right) \\
&= \sum_{n=0}^{\infty} \mathbf{A}^n \left(\sum_{k=0}^{n} \frac{t^k s^{n-k}}{k!\,(n-k)!}\right) \\
&= \sum_{n=0}^{\infty} \mathbf{A}^n \frac{(s+t)^n}{n!} = e^{\mathbf{A}(s+t)}
\end{aligned}$$

Hence

$$e^{\mathbf{A}t}e^{\mathbf{A}s} = e^{\mathbf{A}(t+s)}$$

In particular, if $s = -t$, then

$$e^{\mathbf{A}t}e^{-\mathbf{A}t} = e^{-\mathbf{A}t}e^{\mathbf{A}t} = e^{\mathbf{A}(t-t)} = \mathbf{I}$$

Therefore, the inverse of $e^{\mathbf{A}t}$ is $e^{-\mathbf{A}t}$. Since the inverse exists, $e^{\mathbf{A}t}$ is nonsingular. In general $e^{(\mathbf{A}+\mathbf{B})t}$ and $e^{\mathbf{A}t}e^{\mathbf{B}t}$ are not equal unless $\mathbf{A}$ and $\mathbf{B}$ commute, or $\mathbf{AB} = \mathbf{BA}$.

Let us examine the difference between $e^{(\mathbf{A}+\mathbf{B})t}$ and $e^{\mathbf{A}t}e^{\mathbf{B}t}$. We have

$$e^{(\mathbf{A}+\mathbf{B})t} = \mathbf{I} + (\mathbf{A}+\mathbf{B})t + \frac{(\mathbf{A}+\mathbf{B})^2}{2}t^2 + \frac{(\mathbf{A}+\mathbf{B})^3}{3!}t^3 + \cdots$$

and

$$\begin{aligned} e^{\mathbf{A}t}e^{\mathbf{B}t} &= \left(\mathbf{I} + \mathbf{A}t + \frac{\mathbf{A}^2t^2}{2} + \frac{\mathbf{A}^3t^3}{3!} + \cdots\right) \\ &\quad \cdot \left(\mathbf{I} + \mathbf{B}t + \frac{\mathbf{B}^2t^2}{2} + \frac{\mathbf{B}^3t^3}{3!} + \cdots\right) \\ &= \mathbf{I} + (\mathbf{A}+\mathbf{B})t + \frac{\mathbf{A}^2t^2}{2} + \mathbf{AB}t^2 + \frac{\mathbf{B}^2t^2}{2} \\ &\quad + \frac{\mathbf{A}^3t^3}{3!} + \frac{\mathbf{A}^2\mathbf{B}t^3}{2} + \frac{\mathbf{AB}^2t^3}{2} + \frac{\mathbf{B}^3t^3}{3!} + \cdots \end{aligned}$$

Hence

$$\begin{aligned} &e^{(\mathbf{A}+\mathbf{B})t} - e^{\mathbf{A}t}e^{\mathbf{B}t} \\ &= \frac{(\mathbf{A}+\mathbf{B})(\mathbf{A}+\mathbf{B})}{2}t^2 - \frac{\mathbf{A}^2+2\mathbf{AB}+\mathbf{B}^2}{2}t^2 \\ &\quad + \frac{(\mathbf{A}+\mathbf{B})(\mathbf{A}+\mathbf{B})(\mathbf{A}+\mathbf{B})}{3!}t^3 - \frac{\mathbf{A}^3+3\mathbf{A}^2\mathbf{B}+3\mathbf{AB}^2+\mathbf{B}^3}{3!}t^3 \\ &\quad + \cdots \\ &= \frac{\mathbf{BA}-\mathbf{AB}}{2}t^2 \\ &\quad + \frac{\mathbf{BA}^2+\mathbf{ABA}+\mathbf{B}^2\mathbf{A}+\mathbf{BAB}-2\mathbf{A}^2\mathbf{B}-2\mathbf{AB}^2}{3!}t^3 \\ &\quad + \cdots \end{aligned}$$

If $\mathbf{A}$ and $\mathbf{B}$ commute, or $\mathbf{AB} = \mathbf{BA}$, we see that

$$e^{(\mathbf{A}+\mathbf{B})t} - e^{\mathbf{A}t}e^{\mathbf{B}t} = \mathbf{0}$$

for all t. Hence,

$$e^{(\mathbf{A}+\mathbf{B})t} = e^{\mathbf{A}t}e^{\mathbf{B}t} \qquad \text{if } \mathbf{A} \text{ and } \mathbf{B} \text{ commute}$$

If $\mathbf{A}$ and $\mathbf{B}$ do not commute, then $e^{(\mathbf{A}+\mathbf{B})t}$ is not equal to $e^{\mathbf{A}t}e^{\mathbf{B}t}$.

As an illustrative example, let us compute $e^{\mathbf{A}t}$ where

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

Since

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{B} + \mathbf{C}$$

where

$$\mathbf{B} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

we see that $\mathbf{B}$ and $\mathbf{C}$ commute, or $\mathbf{BC} = \mathbf{CB}$. We have therefore

$$e^{\mathbf{A}t} = e^{(\mathbf{B}+\mathbf{C})t} = e^{\mathbf{B}t}e^{\mathbf{C}t}$$

$e^{\mathbf{C}t}$ is given by

$$\begin{aligned} e^{\mathbf{C}t} &= e^{\mathbf{I}t} \\ &= \mathbf{I} + \mathbf{I}t + \frac{\mathbf{I}^2t^2}{2} + \cdots + \frac{\mathbf{I}^nt^n}{n!} + \cdots \\ &= e^t\mathbf{I} \end{aligned}$$

$e^{\mathbf{B}t}$ is given by

$$e^{\mathbf{B}t} = \mathbf{I} + \mathbf{B}t + \frac{\mathbf{B}^2t^2}{2} + \cdots + \frac{\mathbf{B}^nt^n}{n!} + \cdots$$

where

$$\mathbf{B}^2 = 3\mathbf{B}, \qquad \mathbf{B}^3 = 3\mathbf{B}^2 = 3^2\mathbf{B}, \qquad \mathbf{B}^4 = 3^2\mathbf{B}^2 = 3^3\mathbf{B}, \cdots$$

Hence

$$\begin{aligned} e^{\mathbf{B}t} &= \mathbf{I} + t\mathbf{B} + \frac{t^2}{2}(3\mathbf{B}) + \frac{t^3}{3!}(3^2\mathbf{B}) + \cdots \\ &= \mathbf{I} + \frac{1}{3}\left[3t + \frac{(3t)^2}{2!} + \frac{(3t)^3}{3!} + \cdots\right]\mathbf{B} \\ &= \mathbf{I} + \frac{1}{3}(e^{3t} - 1)\mathbf{B} \end{aligned}$$

Thus we have

$$\begin{aligned} e^{\mathbf{A}t} &= \left[\mathbf{I} + \frac{1}{3}(e^{3t} - 1)\mathbf{B}\right](e^t\mathbf{I}) \\ &= e^t\mathbf{I} + \frac{1}{3}(e^{4t} - e^t)\mathbf{B} \end{aligned}$$

$$= \begin{bmatrix} \frac{1}{3}(e^{4t} + 2e^{t}) & \frac{1}{3}(e^{4t} - e^{t}) & \frac{1}{3}(e^{4t} - e^{t}) \\ \frac{1}{3}(e^{4t} - e^{t}) & \frac{1}{3}(e^{4t} + 2e^{t}) & \frac{1}{3}(e^{4t} - e^{t}) \\ \frac{1}{3}(e^{4t} - e^{t}) & \frac{1}{3}(e^{4t} - e^{t}) & \frac{1}{3}(e^{4t} + 2e^{t}) \end{bmatrix}$$

The direct computation of $e^{\mathbf{A}t}$ as just demonstrated is practical only for simple cases. We shall discuss alternate methods of computation of $e^{\mathbf{A}t}$ in Section 6–4.

We shall next present a theorem on the exponential function of **A**.

Theorem 5-5 For an $n \times n$ matrix **A** having n distinct eigenvectors, $e^{\mathbf{A}}$ can be written as

$$e^{\mathbf{A}} = \mathbf{PMP}^{-1} \qquad (\mathbf{P} = n \times n \text{ nonsingular matrix})$$

where

$$\mathbf{M} = \begin{bmatrix} e^{\lambda_1} & & & 0 \\ & e^{\lambda_2} & & \\ & & \ddots & \\ 0 & & & e^{\lambda_n} \end{bmatrix}$$

and $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of **A** which are not necessarily distinct.

Proof $e^{\mathbf{A}}$ is given as

$$e^{\mathbf{A}} = \sum_{k=0}^{\infty} \frac{1}{k!} \mathbf{A}^k$$

Since **A** can be written as

$$\mathbf{A} = \mathbf{PDP}^{-1}$$

where

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix}$$

we obtain

$$\mathbf{A}^2 = \mathbf{PDP}^{-1}\mathbf{PDP}^{-1} = \mathbf{PD}^2\mathbf{P}^{-1}$$

Therefore, $\mathbf{A}^k$ can be written as follows:

$$\mathbf{A}^k = \mathbf{PD}^k\mathbf{P}^{-1}$$

where

$$\mathbf{D}^k = \begin{bmatrix} \lambda_1^k & & & 0 \\ & \lambda_2^k & & \\ & & \ddots & \\ 0 & & & \lambda_n^k \end{bmatrix}$$

Thus,

$$\sum_{k=0}^{\infty} \frac{1}{k!} \mathbf{D}^k = \begin{bmatrix} \sum_{k=0}^{\infty} \frac{1}{k!} \lambda_1^k & & & 0 \\ & \sum_{k=0}^{\infty} \frac{1}{k!} \lambda_2^k & & \\ & & \ddots & \\ 0 & & & \sum_{k=0}^{\infty} \frac{1}{k!} \lambda_n^k \end{bmatrix}$$

$$= \begin{bmatrix} e^{\lambda_1} & & & 0 \\ & e^{\lambda_2} & & \\ & & \ddots & \\ 0 & & & e^{\lambda_n} \end{bmatrix}$$

$$= \mathbf{M}$$

Hence we obtain

$$\sum_{k=0}^{\infty} \frac{1}{k!} \mathbf{D}^k = \sum_{k=0}^{\infty} \mathbf{P}^{-1}\left(\frac{1}{k!}\mathbf{A}^k\right)\mathbf{P} = \mathbf{P}^{-1}e^{\mathbf{A}}\mathbf{P} = \mathbf{M}$$

This proves the theorem.

5-3 SYLVESTER'S INTERPOLATION FORMULA

In this section we shall first present the well-known Lagrange's interpolation formula and then give detailed discussions of Sylvester's interpolation formula.

Consider the following polynomial of degree $m - 1$, where $\lambda_1, \lambda_2, \ldots, \lambda_m$ are assumed to be distinct:

$$p_k(\lambda) = \frac{(\lambda - \lambda_1) \cdots (\lambda - \lambda_{k-1})(\lambda - \lambda_{k+1}) \cdots (\lambda - \lambda_m)}{(\lambda_k - \lambda_1) \cdots (\lambda_k - \lambda_{k-1})(\lambda_k - \lambda_{k+1}) \cdots (\lambda_k - \lambda_m)}$$

$$(k = 1, 2, \ldots, m)$$

Notice that

$$p_k(\lambda_i) = 1 \quad \text{if} \quad i = k$$
$$= 0 \quad \text{if} \quad i \neq k$$

Consider next the following polynomial $f(\lambda)$ of degree $m - 1$.

$$f(\lambda) = \sum_{k=1}^{m} f(\lambda_k) p_k(\lambda)$$
$$= \sum_{k=1}^{m} f(\lambda_k) \frac{(\lambda - \lambda_1) \cdots (\lambda - \lambda_{k-1})(\lambda - \lambda_{k+1}) \cdots (\lambda - \lambda_m)}{(\lambda_k - \lambda_1) \cdots (\lambda_k - \lambda_{k-1})(\lambda_k - \lambda_{k+1}) \cdots (\lambda_k - \lambda_m)} \tag{5-3}$$

The polynomial $f(\lambda)$ takes on values $f(\lambda_k)$ at points λ_k. Equation (5-3) is the Lagrange's interpolation formula. The $(m - 1)$th-degree polynomial $f(\lambda)$ is determined from m independent data $f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_m)$.

Notice that the $(m - 1)$th-degree polynomial $f(\lambda)$ passes through m points $f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_m)$. Therefore, $f(\lambda)$ is uniquely determined. Any other representations of the polynomial of degree $m - 1$ can be reduced to the Lagrange polynomial $f(\lambda)$.

We can extend this result to the case of a polynomial $f(\mathbf{A})$ in an $n \times n$ matrix $\mathbf{A}$. Let $f_N(\mathbf{A})$ be a polynomial of degree N. Then $f_N(\mathbf{A})$ can be divided by the minimal polynomial $\phi(\mathbf{A})$, yielding

$$f_N(\mathbf{A}) = g(\mathbf{A})\phi(\mathbf{A}) + \alpha_N(\mathbf{A})$$

where $\alpha_N(\mathbf{A})$ is the remainder. If the minimal polynomial $\phi(\mathbf{A})$ is of degree m, then $\alpha_N(\mathbf{A})$ is a polynomial in $\mathbf{A}$ of degree $m - 1$ or less. Since $\phi(\mathbf{A})$ is the minimal polynomial,

$$\phi(\mathbf{A}) = \mathbf{0}$$

Hence,

$$f_N(\mathbf{A}) = \alpha_N(\mathbf{A}) \tag{5-4}$$

If $f_N(\mathbf{A})$ is a convergent infinite series, which we shall denote $f(\mathbf{A})$, then $\alpha_N(\mathbf{A})$ converges to a unique polynomial $\alpha(\mathbf{A})$ as $N \to \infty$. We shall prove this in the following theorem:

Theorem 5-6 A convergent infinite series $f(\mathbf{A}) = \sum_{k=0}^{\infty} c_k \mathbf{A}^k$ of an $n \times n$ matrix $\mathbf{A}$ can be expressed as a unique polynomial in $\mathbf{A}$ of degree $m - 1$ or less, where m is the degree of the minimal polynomial of $\mathbf{A}$.

Proof Let us define

$$f_N(\mathbf{A}) = \sum_{k=0}^{N} c_k \mathbf{A}^k$$

Since $\mathbf{A}^k$ can be expressed as

$$\mathbf{A}^k = \sum_{i=0}^{m-1} b_{ki} \mathbf{A}^i \qquad (k \geqslant 0)$$

we can write

$$f_N(\mathbf{A}) = \sum_{i=0}^{m-1} \alpha_{Ni} \mathbf{A}^i = \alpha_{N0}\mathbf{I} + \alpha_{N1}\mathbf{A} + \cdots + \alpha_{N,m-1}\mathbf{A}^{m-1}$$

where

$$\alpha_{Ni} = \sum_{k=0}^{N} c_k b_{ki} \qquad (i = 0, 1, 2, \ldots, m - 1;\ N \geqslant 0)$$

Hence

$$\lim_{N\to\infty} f_N(\mathbf{A}) = \lim_{N\to\infty} (\alpha_{N0}\mathbf{I} + \alpha_{N1}\mathbf{A} + \cdots + \alpha_{N,m-1}\mathbf{A}^{m-1})$$

By assumption, the series $\lim_{N\to\infty} f_N(\mathbf{A})$ converges to $f(\mathbf{A})$. Since the matrices $\mathbf{I}, \mathbf{A}, \mathbf{A}^2, \ldots, \mathbf{A}^{m-1}$ are linearly independent, if $f(\mathbf{A})$ is expressed as a sum of $\mathbf{I}, \mathbf{A}, \mathbf{A}^2, \ldots, \mathbf{A}^{m-1}$, the coefficients are uniquely determined. Hence

$$\lim_{N\to\infty} f_N(\mathbf{A}) = f(\mathbf{A})$$
$$= (\lim_{N\to\infty} \alpha_{N0})\mathbf{I} + (\lim_{N\to\infty} \alpha_{N1})\mathbf{A} + \cdots + (\lim_{N\to\infty} \alpha_{N,m-1})\mathbf{A}^{m-1}$$
$$= \alpha_0\mathbf{I} + \alpha_1\mathbf{A} + \cdots + \alpha_{m-1}\mathbf{A}^{m-1}$$

where $$\alpha_i = \lim_{N\to\infty} \alpha_{Ni} \qquad (i = 0, 1, \ldots, m - 1)$$

This proves the theorem.

By the use of this theorem we can write a convergent infinite series $f(\mathbf{A})$ of $\mathbf{A}$ as a polynomial in $\mathbf{A}$ of degree $m - 1$ or less, where m is the degree of the minimal polynomial of $\mathbf{A}$, or

$$f(\mathbf{A}) = \alpha(\mathbf{A}) \tag{5-5}$$

where $\alpha(\mathbf{A})$ is a polynomial in $\mathbf{A}$ of degree $m - 1$ or less. From Eq. (5-4) and Eq. (5-5), we conclude that a polynomial in $\mathbf{A}$ of any degree can be expressed as a polynomial in $\mathbf{A}$ of $(m - 1)$th-degree or less.

Since the following analysis applies to both $f_N(\mathbf{A})$ and $f(\mathbf{A})$, we shall drop the subscript N.

We now derive Sylvester's interpolation formula. We first consider the case where the minimal polynomial of $\mathbf{A}$ does not involve multiple roots. Then we shall treat the case where the minimal polynomial of $\mathbf{A}$ does involve multiple roots.

Case 1: where the minimal polynomial involves only distinct roots
Substituting $\mathbf{A}$ for λ in the polynomial $p_k(\lambda)$ we obtain

$$p_k(\mathbf{A}) = \frac{(\mathbf{A} - \lambda_1\mathbf{I}) \cdots (\mathbf{A} - \lambda_{k-1}\mathbf{I})(\mathbf{A} - \lambda_{k+1}\mathbf{I}) \cdots (\mathbf{A} - \lambda_m\mathbf{I})}{(\lambda_k - \lambda_1) \cdots (\lambda_k - \lambda_{k-1})(\lambda_k - \lambda_{k+1}) \cdots (\lambda_k - \lambda_m)}$$

$p_k(\mathbf{A})$ is a polynomial in $\mathbf{A}$ of $(m - 1)$th-degree. Notice that

$$p_k(\lambda_i\mathbf{I}) = \mathbf{I} \quad \text{if} \quad i = k$$
$$= \mathbf{0} \quad \text{if} \quad i \neq k$$

Hence we can write

$$f(\mathbf{A}) = \sum_{k=1}^{m} f(\lambda_k) p_k(\mathbf{A})$$
$$= \sum_{k=1}^{m} f(\lambda_k) \frac{(\mathbf{A} - \lambda_1\mathbf{I}) \cdots (\mathbf{A} - \lambda_{k-1}\mathbf{I})(\mathbf{A} - \lambda_{k+1}\mathbf{I}) \cdots (\mathbf{A} - \lambda_m\mathbf{I})}{(\lambda_k - \lambda_1) \cdots (\lambda_k - \lambda_{k-1})(\lambda_k - \lambda_{k+1}) \cdots (\lambda_k - \lambda_m)} \tag{5-6}$$

Equation (5-6) is known as Sylvester's interpolation formula. Equation (5-6) is equivalent to the following equation:

$$\begin{vmatrix} 1 & 1 & \cdots & 1 & \mathbf{I} \\ \lambda_1 & \lambda_2 & \cdots & \lambda_m & \mathbf{A} \\ \lambda_1^2 & \lambda_2^2 & \cdots & \lambda_m^2 & \mathbf{A}^2 \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \lambda_1^{m-1} & \lambda_2^{m-1} & \cdots & \lambda_m^{m-1} & \mathbf{A}^{m-1} \\ f(\lambda_1) & f(\lambda_2) & \cdots & f(\lambda_m) & f(\mathbf{A}) \end{vmatrix} = 0 \tag{5-7}$$

[Solving Eq. (5-7) for $f(\mathbf{A})$ we obtain Eq. (5-6). Refer to Prob. A-5-4.] Equation (5-7) can also be written as

$$\begin{vmatrix} 1 & \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^{m-1} & f(\lambda_1) \\ 1 & \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^{m-1} & f(\lambda_2) \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ 1 & \lambda_m & \lambda_m^2 & \cdots & \lambda_m^{m-1} & f(\lambda_m) \\ \mathbf{I} & \mathbf{A} & \mathbf{A}^2 & \cdots & \mathbf{A}^{m-1} & f(\mathbf{A}) \end{vmatrix} = 0 \tag{5-8}$$

Notice that solving Eq. (5-8) for $f(\mathbf{A})$ is the same as solving the following set of m linear equations for the α_i $(i = 0, 1, 2, \ldots, m-1)$

$$\begin{aligned} \alpha_0 + \alpha_1\lambda_1 + \alpha_2\lambda_1^2 + \cdots + \alpha_{m-1}\lambda_1^{m-1} &= f(\lambda_1) \\ \alpha_0 + \alpha_1\lambda_2 + \alpha_2\lambda_2^2 + \cdots + \alpha_{m-1}\lambda_2^{m-1} &= f(\lambda_2) \\ \cdots \quad \cdots \quad \cdots & \\ \alpha_0 + \alpha_1\lambda_m + \alpha_2\lambda_m^2 + \cdots + \alpha_{m-1}\lambda_m^{m-1} &= f(\lambda_m) \end{aligned} \tag{5-9}$$

and obtaining $f(\mathbf{A})$ as

$$f(\mathbf{A}) = \alpha_0\mathbf{I} + \alpha_1\mathbf{A} + \alpha_2\mathbf{A}^2 + \cdots + \alpha_{m-1}\mathbf{A}^{m-1} \tag{5-10}$$

Notice also that Eq. (5-10) is equivalent to Eq. (5-4) or Eq. (5-5). Therefore, Eq. (5-9) and Eq. (5-10) can be written as

$$f(\lambda_i) = \alpha(\lambda_i) \qquad (i = 1, 2, \ldots, m)$$
$$f(\mathbf{A}) = \alpha(\mathbf{A})$$

or

$$f(\lambda) = g(\lambda)\phi(\lambda) + \alpha(\lambda), \qquad \phi(\lambda) = \text{minimal polynomial}$$
$$(\lambda = \lambda_1, \lambda_2, \ldots, \lambda_m, \mathbf{A})$$

Case 2: where the minimal polynomial involves multiple roots We shall next consider the case where the minimal polynomial $\phi(\lambda)$ has multiple roots. As a simple example, let us assume that $\phi(\lambda)$ involves three equal roots; that is, the roots of $\phi(\lambda)$ are $\lambda_1, \lambda_1, \lambda_1, \lambda_4, \lambda_5, \ldots, \lambda_m$. Then we may write

$$\phi(\lambda) = \lambda^m + a_1\lambda^{m-1} + \cdots + a_{m-1}\lambda + a_m$$
$$= (\lambda - \lambda_1)^3(\lambda - \lambda_4)(\lambda - \lambda_5) \cdots (\lambda - \lambda_m)$$

Similar to Case 1 we obtain

$$f(\lambda) = g(\lambda)\phi(\lambda) + \alpha(\lambda)$$
$$= g(\lambda)[(\lambda - \lambda_1)^3(\lambda - \lambda_4) \cdots (\lambda - \lambda_m)] + \alpha(\lambda) \tag{5-11}$$

By substituting $\lambda = \lambda_1, \lambda_4, \ldots, \lambda_m$ into Eq. (5-11), we obtain the following $m - 2$ equations:

$$\begin{aligned} f(\lambda_1) &= \alpha(\lambda_1) \\ f(\lambda_4) &= \alpha(\lambda_4) \\ \cdots & \\ f(\lambda_m) &= \alpha(\lambda_m) \end{aligned} \tag{5-12}$$

By differentiating Eq. (5-11) with respect to λ, we obtain

$$\frac{d}{d\lambda} f(\lambda) = (\lambda - \lambda_1)^2 h(\lambda) + \frac{d}{d\lambda}\alpha(\lambda) \tag{5-13}$$

where

$$(\lambda - \lambda_1)^2 h(\lambda) = \frac{d}{d\lambda}[g(\lambda)(\lambda - \lambda_1)^3(\lambda - \lambda_4) \cdots (\lambda - \lambda_m)]$$

Substitution of $\lambda = \lambda_1$ into Eq. (5-13) results in

$$\left.\frac{d}{d\lambda} f(\lambda)\right|_{\lambda=\lambda_1} = f'(\lambda_1) = \left.\frac{d}{d\lambda}\alpha(\lambda)\right|_{\lambda=\lambda_1} \tag{5-14}$$

Similarly, differentiating Eq. (5-11) twice with respect to λ and substituting $\lambda = \lambda_1$, we obtain

$$\left.\frac{d^2}{d\lambda^2} f(\lambda)\right|_{\lambda=\lambda_1} = f''(\lambda_1) = \left.\frac{d^2}{d\lambda^2}\alpha(\lambda)\right|_{\lambda=\lambda_1} \tag{5-15}$$

Rewriting Eq. (5-12), Eq. (5-14), and Eq. (5-15) we have

$$\begin{aligned} \alpha_2 + 3\alpha_3\lambda_1 + \cdots + \frac{(m-1)(m-2)}{2}\alpha_{m-1}\lambda_1^{m-3} &= \frac{f''(\lambda_1)}{2} \\ \alpha_1 + 2\alpha_2\lambda_1 + 3\alpha_3\lambda_1^2 + \cdots + (m-1)\alpha_{m-1}\lambda_1^{m-2} &= f'(\lambda_1) \\ \alpha_0 + \alpha_1\lambda_1 + \alpha_2\lambda_1^2 + \cdots + \alpha_{m-1}\lambda_1^{m-1} &= f(\lambda_1) \\ \alpha_0 + \alpha_1\lambda_4 + \alpha_2\lambda_4^2 + \cdots + \alpha_{m-1}\lambda_4^{m-1} &= f(\lambda_4) \\ \cdots \quad \cdots \quad \cdots & \\ \alpha_0 + \alpha_1\lambda_m + \alpha_2\lambda_m^2 + \cdots + \alpha_{m-1}\lambda_m^{m-1} &= f(\lambda_m) \end{aligned} \tag{5-16}$$

From Eq. (5-16) we can determine the $\alpha_k (k = 0, 1, 2, \ldots, m-1)$ in terms of $f(\lambda_i), f'(\lambda_1)$, and $f''(\lambda_1)$. Then $f(\mathbf{A})$ can be given by

$$f(\mathbf{A}) = \alpha_0\mathbf{I} + \alpha_1\mathbf{A} + \alpha_2\mathbf{A}^2 + \cdots + \alpha_{m-1}\mathbf{A}^{m-1} \tag{5-17}$$

Equations (5–16) and Eq. (5–17) can, of course, be combined in the following way:

$$\begin{vmatrix} 0 & 0 & 1 & 3\lambda_1 & \cdots & \frac{(m-1)(m-2)}{2}\lambda_1^{m-3} & \frac{f''(\lambda_1)}{2} \\ 0 & 1 & 2\lambda_1 & 3\lambda_1^2 & \cdots & (m-1)\lambda_1^{m-2} & f'(\lambda_1) \\ 1 & \lambda_1 & \lambda_1^2 & \lambda_1^3 & \cdots & \lambda_1^{m-1} & f(\lambda_1) \\ 1 & \lambda_4 & \lambda_4^2 & \lambda_4^3 & \cdots & \lambda_4^{m-1} & f(\lambda_4) \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ 1 & \lambda_m & \lambda_m^2 & \lambda_m^3 & \cdots & \lambda_m^{m-1} & f(\lambda_m) \\ \mathbf{I} & \mathbf{A} & \mathbf{A}^2 & \mathbf{A}^3 & \cdots & \mathbf{A}^{m-1} & f(\mathbf{A}) \end{vmatrix} = 0 \qquad (5\text{–}18)$$

If the minimal polynomial has j multiple roots; that is, the roots of $\phi(\lambda)$ are $\lambda_1, \lambda_1, \ldots, \lambda_1, \lambda_{j+1}, \lambda_{j+2}, \ldots, \lambda_m$, then $f(\mathbf{A})$ can be given by

$$f(\mathbf{A}) = \alpha_0\mathbf{I} + \alpha_1\mathbf{A} + \alpha_2\mathbf{A}^2 + \cdots + \alpha_{m-1}\mathbf{A}^{m-1}$$

where the α_k $(k = 0, 1, 2, \ldots, m-1)$ are determined by solving for the coefficients α_k of the following m equations:

$$\begin{gathered} \frac{d^s}{d\lambda_1^s}(\alpha_0 + \alpha_1\lambda_1 + \alpha_2\lambda_1^2 + \cdots + \alpha_{m-1}\lambda_1^{m-1}) = f^{(s)}(\lambda_1) \\ \alpha_0 + \alpha_1\lambda_i + \alpha_2\lambda_i^2 + \cdots + \alpha_{m-1}\lambda_i^{m-1} = f(\lambda_i) \\ (s = 1, 2, \ldots, j-1;\ i = j, j+1, \ldots, m) \end{gathered} \qquad (5\text{–}19)$$

[Equations (5–19) can be solved for the α_k $(k = 0, 1, 2, \ldots, m-1)$, since the determinant of the coefficients is always different from zero. Notice that in this case again the number of the α_k to be determined is m, which is the degree of the minimal polynomial.] Extensions to other cases where the minimal polynomial involves more than two sets of multiple roots are obvious. (The differentiating process is repeated for each set of multiple roots.)

Sylvester's interpolation formula is useful in reducing the exponential function of A into a square matrix. We shall present in Section 6–4 a computational technique for the exponential function of A based on Sylvester's interpolation formula.

5-4 QUADRATIC FORMS AND HERMITIAN FORMS

Quadratic forms appear frequently in problems of stability analysis of dynamic systems by the second method of Liapunov and in problems of optimization, that is, maximizing or minimizing values of functions of variables.

Quadratic forms $\mathbf{x}'\mathbf{A}\mathbf{x}$ may be considered as a special case of more general Hermitian forms $\mathbf{x}^*\mathbf{A}\mathbf{x}$. In fact if A is real symmetric and the variables are constrained to be real, then the Hermitian form $\mathbf{x}^*\mathbf{A}\mathbf{x}$ becomes the quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$. (If the matrix A is real symmetric and the variables are complex, then $\mathbf{x}^*\mathbf{A}\mathbf{x}$ is called a *real Hermitian form*.)

Since in the field of control systems quadratic forms appear more frequently than Hermitian forms, we shall, in this section and the next, devote much attention to quadratic forms. Notice that results obtained for quadratic forms can easily be extended to Hermitian forms.

We shall here give two definitions which we shall use in the subsequent treatment of the subjects:

Rank of quadratic form (or Hermitian form) A quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ (or a Hermitian form $\mathbf{x}^*\mathbf{A}\mathbf{x}$) is called of rank r if A is of rank r.

Diagonal quadratic form (or diagonal Hermitian form) A quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ (or a Hermitian form $\mathbf{x}^*\mathbf{A}\mathbf{x}$) is called a diagonal quadratic form (or a diagonal Hermitian form) if A is a diagonal matrix.

The subjects to be discussed in this section are diagonalization of real symmetric matrices, Sylvester's law of inertia, definiteness, semidefiniteness, and indefiniteness of real quadratic forms and Hermitian forms, and simultaneous reduction of real quadratic forms to diagonal quadratic forms.

We shall now consider a theorem on diagonalization of real symmetric matrices.

Theorem 5-7 Every $n \times n$ real symmetric matrix A can be expressed as

$$\mathbf{A} = \mathbf{T}\mathbf{D}\mathbf{T}' = \mathbf{T}\mathbf{D}\mathbf{T}^{-1}$$

where

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \cdot & \\ & & & \cdot \\ 0 & & & \lambda_n \end{bmatrix}$$

and T is an orthogonal matrix. Also $\mathbf{A}^k$ can be expressed as

$$\mathbf{A}^k = \mathbf{T}\mathbf{D}^k\mathbf{T}' = \mathbf{T}\mathbf{D}^k\mathbf{T}^{-1}$$

where $k = 1, 2, 3, \ldots$

Proof A proof of the first half of this theorem was given in Prob. A-3-3. Here we shall give a more direct proof. Let us define the normalized eigenvectors associated with n eigenvalues of the matrix A to be

$$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$$

namely,

$$(\mathbf{x}_i, \mathbf{x}_i) = 1, \qquad (\mathbf{x}_i, \mathbf{x}_j) = 0 \qquad (i \neq j)$$
$$(i, j = 1, 2, \ldots, n)$$

Let us form the matrix $\mathbf{T}$ whose columns consist of the vectors $\mathbf{x}_i$:

$$\mathbf{T} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix}$$

then

$$\mathbf{T}'\mathbf{T} = \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & \cdot & \\ & & & \cdot \\ & & & \cdot \\ 0 & & & 1 \end{bmatrix}$$

($\mathbf{T}$ is an orthogonal matrix, or $\mathbf{T}' = \mathbf{T}^{-1}$). Notice that since $(\mathbf{A} - \lambda_i \mathbf{I})\mathbf{x}_i = \mathbf{0}$ we obtain

$$\mathbf{AT} = \begin{bmatrix} \lambda_1 x_{11} & \cdots & \lambda_n x_{1n} \\ \lambda_1 x_{21} & \cdots & \lambda_n x_{2n} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \lambda_1 x_{n1} & \cdots & \lambda_n x_{nn} \end{bmatrix} = \mathbf{T} \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \cdot & \\ & & & \cdot \\ & & & \cdot \\ 0 & & & \lambda_n \end{bmatrix}$$

and

$$\mathbf{T}'\mathbf{AT} = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \cdot & \\ & & & \cdot \\ & & & \cdot \\ 0 & & & \lambda_n \end{bmatrix}$$

Therefore

$$\mathbf{A} = \mathbf{T} \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \cdot & \\ & & & \cdot \\ & & & \cdot \\ 0 & & & \lambda_n \end{bmatrix} \mathbf{T}'$$

or

$$\mathbf{A} = \mathbf{TDT}' = \mathbf{TDT}^{-1}$$

Once the first half of the theorem is proved, then the second half can be proved immediately. $\mathbf{A}^2$ can be written as

$$\mathbf{A}^2 = (\mathbf{TDT}')(\mathbf{TDT}') = \mathbf{TDT}'\,\mathbf{TDT}'$$

Since $\mathbf{T}'\mathbf{T} = \mathbf{I}$

$$\mathbf{A}^2 = \mathbf{TD}^2\mathbf{T}' = \mathbf{TD}^2\mathbf{T}^{-1}$$

Similarly,

$$\mathbf{A}^3 = (\mathbf{TD}^2\mathbf{T}')(\mathbf{TDT}') = \mathbf{TD}^3\mathbf{T}' = \mathbf{TD}^3\mathbf{T}^{-1}$$

We thus obtain for $k = 1, 2, 3, \ldots$

$$\mathbf{A}^k = \mathbf{TD}^k\mathbf{T}' = \mathbf{TD}^k\mathbf{T}^{-1}$$

This completes the proof.

Diagonalization of quadratic form From Theorem 5–7 it is clear that by choosing a set of orthonormal eigenvectors of $\mathbf{A}$ as the column vectors of $\mathbf{T}$, the transformation $\mathbf{x} = \mathbf{Ty}$ reduces the quadratic form

$$Q = \mathbf{x}'\mathbf{Ax} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j \qquad (a_{ij} = a_{ji})$$

to

$$Q = \mathbf{y}'\mathbf{T}'\mathbf{ATy} = \sum_{i=1}^{n} \lambda_i y_i^2$$

In the following we shall consider an example.

Example 5-1 Consider the following quadratic form:

$$Q = 2x_1^2 + 2x_2^2 + 4x_3^2 + 2x_1x_2 + 2\sqrt{10}\,x_1x_3 + 2\sqrt{10}\,x_2x_3.$$

Let us find an orthogonal transformation matrix $\mathbf{T}$ that reduces Q to the following form:

$$Q = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \lambda_3 y_3^2$$

where λ_1, λ_2, and λ_3 are the eigenvalues of the square matrix associated with Q.

The quadratic form Q can be written as follows:

$$Q = [x_1\ x_2\ x_3] \begin{bmatrix} 2 & 1 & \sqrt{10} \\ 1 & 2 & \sqrt{10} \\ \sqrt{10} & \sqrt{10} & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$
$$= \mathbf{x}'\mathbf{Ax}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} 2 & 1 & \sqrt{10} \\ 1 & 2 & \sqrt{10} \\ \sqrt{10} & \sqrt{10} & 4 \end{bmatrix}$$

Notice that **A** is a symmetric matrix. The characteristic equation is

$$|\mathbf{A}-\lambda\mathbf{I}|=\begin{vmatrix}2-\lambda & 1 & \sqrt{10}\\ 1 & 2-\lambda & \sqrt{10}\\ \sqrt{10} & \sqrt{10} & 4-\lambda\end{vmatrix}=0$$

The eigenvalues are found to be

$$\lambda_1=-1,\qquad \lambda_2=1,\qquad \lambda_3=8$$

A transformation matrix can be determined from Eq. (3–18) as follows:

$$\begin{bmatrix}5 & -7 & 14\\ 5 & 7 & 14\\ -2\sqrt{10} & 0 & 7\sqrt{10}\end{bmatrix}$$

Each column of this matrix is an eigenvector of **A**. Normalizing the eigenvectors, we obtain the transformation matrix **T** as follows:

$$\mathbf{T}=\begin{bmatrix}\frac{\sqrt{5}}{\sqrt{18}} & -\frac{1}{\sqrt{2}} & \frac{\sqrt{2}}{3}\\ \frac{\sqrt{5}}{\sqrt{18}} & \frac{1}{\sqrt{2}} & \frac{\sqrt{2}}{3}\\ -\frac{2}{3} & 0 & \frac{\sqrt{5}}{3}\end{bmatrix}$$

(Notice that $\mathbf{T}'\mathbf{T}=\mathbf{I}$.) The transformation

$$\mathbf{x}=\mathbf{T}\mathbf{y}$$

reduces

$$Q=\mathbf{x}'\mathbf{A}\mathbf{x}$$

to

$$\begin{aligned}Q&=\mathbf{y}'\mathbf{T}'\mathbf{A}\mathbf{T}\mathbf{y}=\mathbf{y}'\mathbf{D}\mathbf{y}\\ &=[y_1\ y_2\ y_3]\begin{bmatrix}-1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 8\end{bmatrix}\begin{bmatrix}y_1\\ y_2\\ y_3\end{bmatrix}\\ &=-y_1^2+y_2^2+8y_3^2\end{aligned}$$

In other words, by the transformation

$$\begin{aligned}x_1&=\frac{\sqrt{5}}{\sqrt{18}}y_1-\frac{1}{\sqrt{2}}y_2+\frac{\sqrt{2}}{3}y_3\\ x_2&=\frac{\sqrt{5}}{\sqrt{18}}y_1+\frac{1}{\sqrt{2}}y_2+\frac{\sqrt{2}}{3}y_3\\ y_3&=-\frac{2}{3}y_1+\frac{\sqrt{5}}{3}y_3\end{aligned}$$

the quadratic form

$$Q=2x_1^2+2x_2^2+4x_3^2+2x_1x_2+2\sqrt{10}\,x_1x_3+2\sqrt{10}\,x_2x_3$$

can be reduced to

$$Q=-y_1^2+y_2^2+8y_3^2$$

Diagonalization of Hermitian form Diagonalization of Hermitian forms can be carried out by essentially the same technique as that employed for quadratic forms. Since only minor modifications are necessary, we do not discuss it in detail, but merely give an illustrative example.

Example 5-2 Consider the following Hermitian form:

$$\begin{aligned}H=&\bar{x}_1x_1+2\bar{x}_2x_2+\bar{x}_3x_3\\ &+(1+j\sqrt{2})\bar{x}_1x_2+(1-j\sqrt{2})\bar{x}_2x_1\\ &+j\sqrt{3}\,\bar{x}_2x_3-j\sqrt{3}\,\bar{x}_3x_2\end{aligned}$$

Let us find a unitary transformation **U** that converts H into the following form:

$$H=\lambda_1\bar{y}_1y_1+\lambda_2\bar{y}_2y_2+\lambda_3\bar{y}_3y_3$$

where λ_1, λ_2, and λ_3 are the eigenvalues of the square matrix associated with H.

The given Hermitian form can be written as

$$H=[\bar{x}_1\ \bar{x}_2\ \bar{x}_3]\begin{bmatrix}1 & 1+j\sqrt{2} & 0\\ 1-j\sqrt{2} & 2 & j\sqrt{3}\\ 0 & -j\sqrt{3} & 1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}$$

Referring to Example 3–1, the eigenvalues of **A** are

$$\lambda_1=1,\qquad \lambda_2=-1,\qquad \lambda_3=4$$

The unitary matrix **U** whose columns consist of the normalized eigenvectors of **A** was obtained in Example 3–1 as

$$\mathbf{U}=\begin{bmatrix}\frac{-j\sqrt{3}}{\sqrt{6}} & \frac{1+j\sqrt{2}}{\sqrt{10}} & \frac{1+j\sqrt{2}}{\sqrt{15}}\\ 0 & \frac{-2}{\sqrt{10}} & \frac{3}{\sqrt{15}}\\ \frac{1-j\sqrt{2}}{\sqrt{6}} & \frac{-j\sqrt{3}}{\sqrt{10}} & \frac{-j\sqrt{3}}{\sqrt{15}}\end{bmatrix}$$

The transformation $\mathbf{x}=\mathbf{U}\mathbf{y}$ will transform $H=\mathbf{x}^*\mathbf{A}\mathbf{x}$ into

$$H=\mathbf{y}^*\mathbf{U}^*\mathbf{A}\mathbf{U}\mathbf{y}=\mathbf{y}^*\mathbf{D}\mathbf{y}=[\bar{y}_1\ \bar{y}_2\ \bar{y}_3]\begin{bmatrix}1 & 0 & 0\\ 0 & -1 & 0\\ 0 & 0 & 4\end{bmatrix}\begin{bmatrix}y_1\\ y_2\\ y_3\end{bmatrix}$$

In other words, by the transformation,

$$x_1=\frac{-j\sqrt{3}}{\sqrt{6}}y_1+\frac{1+j\sqrt{2}}{\sqrt{10}}y_2+\frac{1+j\sqrt{2}}{\sqrt{15}}y_3$$

$$x_2 = -\frac{2}{\sqrt{10}}y_2 + \frac{3}{\sqrt{15}}y_3$$

$$x_3 = \frac{1-j\sqrt{2}}{\sqrt{6}}y_1 - \frac{j\sqrt{3}}{\sqrt{10}}y_2 - \frac{j\sqrt{3}}{\sqrt{15}}y_3$$

the Hermitian form

$$\begin{aligned} H = {} & \bar{x}_1 x_1 + 2\bar{x}_2 x_2 + \bar{x}_3 x_3 \\ & + (1 + j\sqrt{2})\,\bar{x}_1 x_2 + (1 - j\sqrt{2})\,\bar{x}_2 x_1 \\ & + j\sqrt{3}\,\bar{x}_2 x_3 - j\sqrt{3}\,\bar{x}_3 x_2 \end{aligned}$$

can be transformed into

$$H = \bar{y}_1 y_1 - \bar{y}_2 y_2 + 4\bar{y}_3 y_3$$

Theorem 5-8 (*Sylvester's law of inertia*) If a quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ of rank m is reduced by a nonsingular transformation $\mathbf{x} = \mathbf{T}\mathbf{y}$ to the diagonal form

$$a_1 y_1^2 + a_2 y_2^2 + \cdots + a_p y_p^2 + a_{p+1} y_{p+1}^2 + \cdots + a_{p+q} y_{p+q}^2$$

then the number p of positive coefficients $a_1, a_2, \ldots, a_p$ and the number q of negative coefficients $a_{p+1}, a_{p+2}, \ldots, a_{p+q}$ are uniquely determined by $\mathbf{A}$. Moreover, p is equal to the number of positive eigenvalues of $\mathbf{A}$ and q is equal to the number of negative eigenvalues of $\mathbf{A}$.

Proof Let us define

$$Q_1 = a_1 y_1^2 + \cdots + a_p y_p^2 + a_{p+1} y_{p+1}^2 + \cdots + a_{p+q} y_{p+q}^2 \tag{5-20}$$

Consider a nonsingular transformation $\mathbf{x} = \mathbf{S}\mathbf{z}$ which reduces $\mathbf{x}'\mathbf{A}\mathbf{x}$ into

$$Q_2 = b_1 z_1^2 + \cdots + b_r z_r^2 + b_{r+1} z_{r+1}^2 + \cdots + b_{r+s} z_{r+s}^2 \tag{5-21}$$

where $b_1, b_2, \ldots, b_r$ are positive and $b_{r+1}, b_{r+2}, \ldots, b_{r+s}$ are negative. We shall first prove that $p = r$ and $q = s$.

Since $p + q = \text{rank } \mathbf{A} = m$, we have $p + q = r + s$. Since the two forms Eq. (5-20) and Eq. (5-21) are obtained from the same $\mathbf{x}'\mathbf{A}\mathbf{x}$, we have

$$\sum_{i=1}^{p} a_i y_i^2 + \sum_{i=p+1}^{p+q} a_i y_i^2 = \sum_{i=1}^{r} b_i z_i^2 + \sum_{i=r+1}^{r+s} b_i z_i^2 \tag{5-22}$$

Since $\mathbf{y}$ and $\mathbf{z}$ are related to $\mathbf{x}$ by

$$\mathbf{y} = \mathbf{T}^{-1}\mathbf{x}$$

$$\mathbf{z} = \mathbf{S}^{-1}\mathbf{x}$$

respectively, $y_1, y_2, \ldots, y_n$ and $z_1, z_2, \ldots, z_n$ can be expressed as linear functions of $x_1, x_2, \ldots, x_n$.

Suppose that $p > r$. Then a set of $n - p + r$ linear homogeneous equations

$$\begin{aligned} y_i &= 0 \qquad (p+1 \leqslant i \leqslant n) \\ z_j &= 0 \qquad (1 \leqslant j \leqslant r) \end{aligned}$$

in the variables $x_1, x_2, \ldots, x_n$ possesses a nontrivial solution

$$\mathbf{x} = \mathbf{x}_0$$

This is because

$$\begin{aligned} \text{number of equations} &= n - p + r \\ &< n \\ &= \text{number of variables} \end{aligned}$$

Define

$$\mathbf{y} = \mathbf{T}^{-1}\mathbf{x}_0 = \begin{bmatrix} y_{10} \\ y_{20} \\ \vdots \\ y_{p0} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad \mathbf{z} = \mathbf{S}^{-1}\mathbf{x}_0 = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ z_{r+1,0} \\ z_{r+2,0} \\ \vdots \\ z_{n0} \end{bmatrix}$$

From Eq. (5-22) we have

$$\sum_{i=1}^{p} a_i y_i^2 - \sum_{i=r+1}^{r+s} b_i z_i^2 = \sum_{i=1}^{r} b_i z_i^2 - \sum_{i=p+1}^{p+q} a_i y_i^2$$

But the right-hand side of this equation is equal to zero, since $z_i = 0$ $(i = 1, 2, \ldots, r)$ and $y_j = 0$ $(j = p+1, p+2, \ldots, p+q)$. This is a contradiction, since

$$\begin{bmatrix} y_{10} \\ \vdots \\ y_{p0} \end{bmatrix} \neq \mathbf{0}, \qquad \begin{bmatrix} z_{r+1,0} \\ \vdots \\ z_{n0} \end{bmatrix} \neq \mathbf{0}$$

Therefore, $p \leqslant r$. Similarly, we can prove that $r \leqslant p$. Hence $p = r$. Since $p + q = r + s$ it follows that $q = s$. Hence we have proved that

$$p = r \qquad \text{and} \qquad q = s$$

We shall next prove that p is equal to the number of positive eigenvalues of $\mathbf{A}$ and q is equal to the number of negative eigenvalues of $\mathbf{A}$. Since $\mathbf{A}$ has

rank m there are m nonzero eigenvalues. Let us define positive eigenvalues as $\lambda_1, \lambda_2, \ldots, \lambda_h$ and negative eigenvalues as $\lambda_{h+1}, \lambda_{h+2}, \ldots, \lambda_{h+k}$ $(h + k = m)$. By Theorem 5–7 there exists an orthogonal matrix $\mathbf{T}$ such that the transformation $\mathbf{x} = \mathbf{T}\boldsymbol{\xi}$ will reduce $\mathbf{x}'\mathbf{A}\mathbf{x}$ into the following diagonal quadratic form:

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \boldsymbol{\xi}'\mathbf{T}'\mathbf{A}\mathbf{T}\boldsymbol{\xi} = \boldsymbol{\xi}' \begin{bmatrix} \lambda_1 & & & & & 0 \\ & \ddots & & & & \\ & & \lambda_m & & & \\ & & & 0 & & \\ & & & & \ddots & \\ 0 & & & & & 0 \end{bmatrix} \boldsymbol{\xi}$$

Hence

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \lambda_1\xi_1^2 + \cdots + \lambda_h\xi_h^2 + \lambda_{h+1}\xi_{h+1}^2 + \cdots + \lambda_{h+k}\xi_{h+k}^2$$
$$(h + k = m)$$

Because of the uniqueness of p and q which we have proved, we obtain

$$p = r = h \qquad \text{and} \qquad q = s = k$$

This completes the proof.

Notice that the reduction to the diagonal form is not unique. The elements in the diagonal matrix $\mathbf{T}'\mathbf{A}\mathbf{T}$ are not unique, except that the number of positive elements and that of negative elements are uniquely determined by $\mathbf{A}$.

Theorem 5-9 A quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ of rank m can be reduced by a real nonsingular transformation $\mathbf{x} = \mathbf{P}\mathbf{y}$ to the diagonal form

$$y_1^2 + \cdots + y_p^2 - y_{p+1}^2 - \cdots - y_{p+q}^2$$

where the numbers p and q are, respectively, the number of positive eigenvalues and that of negative eigenvalues of $\mathbf{A}$; and $p + q = m$.

Proof Since $\mathbf{A}$ is a real symmetric matrix, the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are real. Let us assume that eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ are positive real, $\lambda_{p+1}, \lambda_{p+2}, \ldots, \lambda_{p+q}$ are negative real and $\lambda_{p+q+1}, \lambda_{p+q+2}, \ldots, \lambda_n$ are zero. Let $\mathbf{T}$ be an orthogonal matrix such that $\mathbf{T}'\mathbf{A}\mathbf{T} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Let us define $\mathbf{P}$ by

$$\mathbf{P} = \mathbf{T}\mathbf{D} = \mathbf{T} \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} & & & & & & 0 \\ & \ddots & & & & & \\ & & \frac{1}{\sqrt{\lambda_p}} & & & & \\ & & & \frac{1}{\sqrt{-\lambda_{p+1}}} & & & \\ & & & & \ddots & & \\ & & & & & \frac{1}{\sqrt{-\lambda_{p+q}}} & \\ & & & & & & 1 \\ & & & & & & & \ddots \\ 0 & & & & & & & & 1 \end{bmatrix}$$

then

$$\mathbf{P}'\mathbf{A}\mathbf{P} = \mathbf{D}'\mathbf{T}'\mathbf{A}\mathbf{T}\mathbf{D} = \mathbf{D}' \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix} \mathbf{D}$$

$$= \left[\begin{array}{ccc:ccc:c} 1 & & 0 & & & & 0 \\ & \ddots & & & & & \\ 0 & & 1 & & & & \\ \hdashline & & & -1 & & 0 & \\ & & & & \ddots & & \\ & & & 0 & & -1 & \\ \hdashline & & & & & & 0 \\ & & & & & & \\ 0 & & & & & & 0 \end{array}\right] \begin{matrix} \\ \\ \}p \\ \\ \\ \}q \\ \\ \\ \end{matrix}$$

Hence we obtain

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{y}'\,\mathbf{P}'\mathbf{A}\mathbf{P}\mathbf{y}$$
$$= y_1^2 + \cdots + y_p^2 - y_{p+1}^2 - \cdots - y_{p+q}^2$$

This proves the theorem.

Signature of quadratic form x′Ax (or Hermitian form x*Ax) The number $p - q$ is called the *signature* of the quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ (or Hermitian form $\mathbf{x}^*\mathbf{A}\mathbf{x}$) or of the matrix $\mathbf{A}$.

Notice that for two real symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ to be similar, it is necessary and sufficient that the ranks of $\mathbf{A}$ and $\mathbf{B}$ be the same and the signatures of $\mathbf{A}$ and $\mathbf{B}$ be the same.

Definiteness, semidefiniteness, and indefiniteness of quadratic form The quadratic form

$$y_1^2 + \cdots + y_p^2 - y_{p+1}^2 - \cdots - y_{p+q}^2 + 0y_{p+q+1}^2 + \cdots + 0y_n^2$$

is

1 Positive definite if $p = n, q = 0$
2 Positive semidefinite if $p < n, q = 0$
3 Negative definite if $p = 0, q = n$
4 Negative semidefinite if $p = 0, q < n$
5 Indefinite if $pq \neq 0$

Definiteness, semidefiniteness and indefiniteness of Hermitian form Since the Hermitian form yields real values, it can also be classified as positive (negative) definite, positive (negative) semidefinite, and indefinite. A Hermitian form

$$\bar{y}_1 y_1 + \cdots + \bar{y}_p y_p - \bar{y}_{p+1} y_{p+1} - \cdots - \bar{y}_{p+q} y_{p+q} + 0\bar{y}_{p+q+1} y_{p+q+1} + \cdots + 0\bar{y}_n y_n$$

is

1 Positive definite if $p = n, q = 0$
2 Positive semidefinite if $p < n, q = 0$
3 Negative definite if $p = 0, q = n$
4 Negative semidefinite if $p = 0, q < n$
5 Indefinite if $pq \neq 0$

A real symmetric matrix $\mathbf{A}$ or a Hermitian matrix $\mathbf{A}$ is called positive (negative) definite, positive (negative) semidefinite, or indefinite, if the corresponding real quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ or the corresponding Hermitian form $\mathbf{x}^*\mathbf{A}\mathbf{x}$ is positive (negative) definite, positive (negative) semidefinite, or indefinite, respectively.

Notice that a real symmetric matrix $\mathbf{A}$ or Hermitian matrix $\mathbf{A}$ is positive definite implies that $\lambda_1 > 0, \lambda_2 > 0, \ldots, \lambda_n > 0$. Therefore, if $\mathbf{A}$ is positive definite, then

$$|\mathbf{A}| = \prod_{i=1}^{n} \lambda_i > 0$$

Thus we find that the determinant of a positive definite matrix $\mathbf{A}$ is positive. Similarly, we find the determinant of a negative definite matrix $\mathbf{A}$ is negative if n is odd, positive if n is even; and the determinant of a semi-definite matrix is zero. Hence a positive or negative definite matrix $\mathbf{A}$ always has an inverse.

Notice that if λ_i is an eigenvalue of $\mathbf{A}$, then λ_i^{-1} is an eigenvalue of $\mathbf{A}^{-1}$. [Refer to Eq. (2–6) on page 77.] Hence $\mathbf{A}$ being positive (negative) definite implies that $\mathbf{A}^{-1}$ being positive (negative) definite.

Notice also that if $\mathbf{x}'\mathbf{A}\mathbf{x}$ (or $\mathbf{x}^*\mathbf{A}\mathbf{x}$) is positive definite and $\mathbf{x} = \mathbf{T}\mathbf{y}$ is a nonsingular transformation, then $\mathbf{y}'\mathbf{T}'\mathbf{A}\mathbf{T}\mathbf{y}$ (or $\mathbf{y}^*\mathbf{T}^*\mathbf{A}\mathbf{T}\mathbf{y}$) is also positive definite. Therefore, $\mathbf{A}$ being positive definite implies $\mathbf{T}'\mathbf{A}\mathbf{T}$ (or $\mathbf{T}^*\mathbf{A}\mathbf{T}$) being positive definite and vice versa.

We shall next present a theorem on Hermitian forms.

Theorem 5-10 If a Hermitian form $\mathbf{x}^*\mathbf{A}\mathbf{x}$ of rank m is reduced by a nonsingular transformation $\mathbf{x} = \mathbf{T}\mathbf{y}$ to the diagonal form

$$a_1\bar{y}_1 y_1 + \cdots + a_p\bar{y}_p y_p + a_{p+1}\bar{y}_{p+1} y_{p+1} + \cdots + a_{p+q}\bar{y}_{p+q} y_{p+q}$$

then the number p of positive coefficients $a_1, a_2, \ldots, a_p$ and the number q of negative coefficients $a_{p+1}, a_{p+2}, \ldots, a_{p+q}$ are uniquely determined by $\mathbf{A}$. Moreover, p is equal to the number of positive eigenvalues of $\mathbf{A}$ and q is equal to the number of negative eigenvalues of $\mathbf{A}$, and thus $p + q = m$.

Proof We shall leave the proof to the reader, since it is quite similar to that of Theorem 5–8.

We shall next present a theorem on positive definite real symmetric matrices.

Theorem 5-11 If $\mathbf{A}$ is an $n \times n$ positive definite real symmetric matrix, then there exists a nonsingular matrix $\mathbf{S}$ such that

$$\mathbf{A} = \mathbf{S}\mathbf{S}'$$

Proof It has been proved that there exists an orthogonal matrix $\mathbf{T}$ such that

$$\mathbf{T}'\mathbf{A}\mathbf{T} = \mathbf{D} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

Since $\mathbf{A}$ is assumed to be real symmetric and positive definite, all the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are positive. Hence

$$\text{diag}\,(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n})$$

is a real diagonal matrix. We can therefore write

$$\mathbf{A} = \mathbf{TDT}' = \mathbf{T}\begin{bmatrix} \sqrt{\lambda_1} & & & 0 \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ 0 & & & \sqrt{\lambda_n} \end{bmatrix}\begin{bmatrix} \sqrt{\lambda_1} & & & 0 \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ 0 & & & \sqrt{\lambda_n} \end{bmatrix}\mathbf{T}'$$
$$= \mathbf{SS}'$$

where $$\mathbf{S} = \mathbf{T}\,\text{diag}\,(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n})$$

This completes the proof.

We shall now consider simultaneous reduction of quadratic forms into diagonal forms.

Theorem 5-12 If $\mathbf{A}$ is an $n \times n$ positive definite real symmetric matrix and $\mathbf{B}$ is an $n \times n$ real symmetric matrix, then there exists a nonsingular matrix $\mathbf{S}$ such that

$$\mathbf{S}'\mathbf{AS} = \mathbf{I}$$

and

$$\mathbf{S}'\mathbf{BS} = \text{diag}\,(\mu_1, \mu_2, \ldots, \mu_n)$$

where $\mu_1, \mu_2, \ldots, \mu_n$ are the eigenvalues of $|\mathbf{B} - \mu\mathbf{A}| = 0$.

Proof It has been proved that there exists an orthogonal matrix $\mathbf{T}$ such that

$$\mathbf{T}'\mathbf{AT} = \mathbf{D} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

where the λ_i are the eigenvalues of $\mathbf{A}$. (The λ_i are all positive real.) Let us define

$$\mathbf{Q} = \mathbf{T}\,\text{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{\lambda_2}}, \ldots, \frac{1}{\sqrt{\lambda_n}}\right)$$

then

$$\mathbf{Q}'\mathbf{AQ} = \text{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \ldots, \frac{1}{\sqrt{\lambda_n}}\right)\mathbf{T}'\mathbf{AT}\,\text{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \ldots, \frac{1}{\sqrt{\lambda_n}}\right)$$
$$= \mathbf{I}$$

Since $\mathbf{B}$ is real symmetric, $\mathbf{Q}'\mathbf{BQ}$ is also a real symmetric matrix. There exists an orthogonal matrix $\mathbf{Z}$ such that $\mathbf{Z}'(\mathbf{Q}'\mathbf{BQ})\mathbf{Z}$ is diagonal. The diagonal elements are the eigenvalues of $\mathbf{Q}'\mathbf{BQ}$ and are the roots of $|\mathbf{B} - \mu\mathbf{A}| = 0$, since

$$\mathbf{Q}'\mathbf{BQ} - \mu\mathbf{I} = \mathbf{Q}'\mathbf{BQ} - \mu\mathbf{Q}'\mathbf{AQ} = \mathbf{Q}'(\mathbf{B} - \mu\mathbf{A})\mathbf{Q}$$

and $|\mathbf{Q}| \neq 0$. Hence

$$\mathbf{Z}'(\mathbf{Q}'\mathbf{BQ})\mathbf{Z} = \text{diag}\,(\mu_1, \mu_2, \ldots, \mu_n)$$

Since the matrix $\mathbf{Z}$ is orthogonal, and $\mathbf{Z}'\mathbf{Z} = \mathbf{I}$ the desired matrix $\mathbf{S}$ is given by $\mathbf{S} = \mathbf{QZ}$. With this matrix $\mathbf{S}$ we obtain

$$\mathbf{S}'\mathbf{AS} = \mathbf{Z}'\mathbf{Q}'\mathbf{AQZ} = \mathbf{Z}'\mathbf{IZ} = \mathbf{I}$$

and

$$\mathbf{S}'\mathbf{BS} = (\mathbf{Z}'\mathbf{Q}')\,\mathbf{B}(\mathbf{QZ}) = \text{diag}\,(\mu_1, \mu_2, \ldots, \mu_n)$$

This proves the theorem.

Notice that the roots of $|\mathbf{B} - \mu\mathbf{A}| = 0$ are all real, since they are also the roots of $|\mathbf{Q}'\,\mathbf{BQ} - \mu\mathbf{I}| = 0$ and $\mathbf{Q}'\,\mathbf{BQ}$ is a real symmetric matrix.

Example 5-3 The kinetic energy T and the potential energy V of a dynamic system are given as follows:

$$T = 5\dot{x}_1^2 - 4\dot{x}_1\dot{x}_2 + 2\dot{x}_2^2$$
$$V = 2x_1^2 - 4x_1x_2 + 6x_2^2$$

By use of a suitable transformation matrix $\mathbf{S}$ reduce the quadratic forms of the kinetic energy and potential energy into diagonal quadratic forms, or

$$T = \dot{y}_1^2 + \dot{y}_2^2$$
$$V = a_1y_1^2 + a_2y_2^2$$

where a_1 and a_2 are constants. Obtain Lagrange's equations of motion in terms of y_1 and y_2.

T and V can be written as

$$T = [\dot{x}_1\ \dot{x}_2]\begin{bmatrix} 5 & -2 \\ -2 & 2 \end{bmatrix}\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \dot{\mathbf{x}}'\mathbf{A}\dot{\mathbf{x}}$$

$$V = [x_1\ x_2]\begin{bmatrix} 2 & -2 \\ -2 & 6 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{x}'\mathbf{Bx}$$

Notice that $\mathbf{A}$ and $\mathbf{B}$ are real symmetric and both are positive definite. The eigenvalues of $\mathbf{A}$ are $\lambda_1 = 1, \lambda_2 = 6$. The eigenvectors of $\mathbf{A}$ are found as

$$\begin{bmatrix} a \\ 2a \end{bmatrix}, \quad \begin{bmatrix} 2b \\ -b \end{bmatrix}$$

where a and b are constants. The orthogonal transformation matrix $\mathbf{T}$ is then obtained as

$$\mathbf{T} = \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} \end{bmatrix}$$

Then $\mathbf{Q}$ is found as

$$\mathbf{Q} = \mathbf{T}\,\mathrm{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{\lambda_2}}\right) = \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{30}} \\ \frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{30}} \end{bmatrix}$$

and $\mathbf{Q}'\mathbf{B}\mathbf{Q}$ becomes

$$\mathbf{Q}'\mathbf{B}\mathbf{Q} = \begin{bmatrix} \frac{18}{5} & -\frac{14}{\sqrt{150}} \\ -\frac{14}{\sqrt{150}} & \frac{11}{15} \end{bmatrix}$$

The eigenvalues of $\mathbf{Q}'\mathbf{B}\mathbf{Q}$ are the roots of $|\mathbf{B} - \mu\mathbf{A}| = 0$. Solving

$$|\mathbf{B} - \mu\mathbf{A}| = \begin{vmatrix} 2 - 5\mu & -2 + 2\mu \\ -2 + 2\mu & 6 - 2\mu \end{vmatrix} = 0$$

we obtain $\mu_1 = \frac{1}{3}$, $\mu_2 = 4$. Once the eigenvalues of $\mathbf{Q}'\mathbf{B}\mathbf{Q}$ are found to be $\frac{1}{3}$ and 4, the transformation matrix $\mathbf{Z}$ is obtained as follows:

$$\mathbf{Z} = \begin{bmatrix} \frac{\sqrt{6}}{\sqrt{55}} & \frac{7}{\sqrt{55}} \\ \frac{7}{\sqrt{55}} & -\frac{\sqrt{6}}{\sqrt{55}} \end{bmatrix}$$

The desired transformation matrix $\mathbf{S}$ is given by

$$\mathbf{S} = \mathbf{Q}\mathbf{Z} = \begin{bmatrix} \frac{4}{\sqrt{66}} & \frac{1}{\sqrt{11}} \\ \frac{1}{\sqrt{66}} & \frac{3}{\sqrt{11}} \end{bmatrix}$$

With the transformation $\mathbf{x} = \mathbf{S}\mathbf{y}$, the quadratic forms T and V are reduced to

$$T = \dot{\mathbf{x}}'\mathbf{A}\dot{\mathbf{x}} = \dot{\mathbf{y}}'\mathbf{S}'\mathbf{A}\mathbf{S}\dot{\mathbf{y}} = \dot{y}_1^2 + \dot{y}_2^2$$
$$V = \mathbf{x}'\mathbf{B}\mathbf{x} = \mathbf{y}'\mathbf{S}'\mathbf{B}\mathbf{S}\mathbf{y} = \tfrac{1}{3}y_1^2 + 4y_2^2$$

The Lagrange's equations of motion are

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{y}_i}\right) - \frac{\partial L}{\partial y_i} = 0 \qquad (i = 1, 2)$$

where $L = T - V$ = Lagrangian

For the given system we obtain

$$\ddot{y}_1 + \tfrac{1}{3}y_1 = 0$$
$$\ddot{y}_2 + 4y_2 = 0$$

which can be solved immediately. The solution in the original coordinates x_1 and x_2 is given by

$$x_1 = \frac{4}{\sqrt{66}}y_1 + \frac{1}{\sqrt{11}}y_2$$
$$x_2 = \frac{1}{\sqrt{66}}y_1 + \frac{3}{\sqrt{11}}y_2$$

5-5 SYLVESTER'S CRITERIA FOR DEFINITENESS AND SEMIDEFINITENESS OF QUADRATIC FORMS AND HERMITIAN FORMS

In this section we shall first prove Sylvester's criterion for positive definiteness of quadratic forms. We shall then present Sylvester's criteria for negative definiteness and for positive or negative semidefiniteness of quadratic forms. Finally we shall state that Sylvester's criteria for definiteness or semidefiniteness of quadratic forms can be applied to Hermitian forms.

Theorem 5-13 (Sylvester's criterion for positive definiteness of quadratic form) A necessary and sufficient condition in order that the quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is an $n \times n$ real symmetric matrix, be positive definite is that the determinant of $\mathbf{A}$ be positive and the successive principal minors of the determinant of $\mathbf{A}$ be positive, or

$$a_{11} > 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0, \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} > 0, \quad \cdots$$

$$\det \mathbf{A} = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} > 0 \qquad (a_{ij} = a_{ji})$$

Proof Assume that the quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ is positive definite. In the following quadratic equation

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j \qquad (a_{ij} = a_{ji})$$

let us put

$$x_{k+1} = x_{k+2} = \cdots = x_n = 0$$

Then

$$\sum_{r=1}^{k}\sum_{s=1}^{k} a_{rs}x_r x_s \qquad (a_{rs} = a_{sr})$$

is also positive definite. The coefficient matrix in this case is given by

$$\mathbf{A}_k = \begin{vmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{kk} \end{vmatrix}$$

If $\mathbf{A}_k$ is positive definite, then $|\mathbf{A}_k| > 0$, or

$$\begin{vmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{kk} \end{vmatrix} > 0$$

Hence, we conclude that if the quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ is positive definite, then the successive principal minors of the determinant of $\mathbf{A}$ are positive.

We shall next prove the converse; namely, we shall prove that if

$$|\mathbf{A}_k| > 0 \qquad (k = 1, 2, \ldots, n)$$

then $\mathbf{x}'\mathbf{A}\mathbf{x}$ is positive definite. We shall use induction. For $k = 1$, it is clear. Assume that the result holds for the $(k-1) \times (k-1)$ matrix $\mathbf{A}_{k-1}$. Let us write the $k \times k$ real symmetric matrix $\mathbf{A}_k$ as

$$\mathbf{A}_k = \left[\begin{array}{c|c} \mathbf{A}_{k-1} & \mathbf{a} \\ \hline \mathbf{a}' & a_{kk} \end{array}\right]$$

which can be rewritten as (refer to Prob. A-5-14):

$$\mathbf{A}_k = \left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline \mathbf{a}'\mathbf{A}_{k-1}^{-1} & 1 \end{array}\right] \left[\begin{array}{c|c} \mathbf{A}_{k-1} & \mathbf{0} \\ \hline \mathbf{0} & a_{kk} - \mathbf{a}'\mathbf{A}_{k-1}^{-1}\mathbf{a} \end{array}\right] \left[\begin{array}{c|c} \mathbf{I} & \mathbf{A}_{k-1}^{-1}\mathbf{a} \\ \hline \mathbf{0} & 1 \end{array}\right]$$

From this we obtain

$$|\mathbf{A}_k| = |\mathbf{A}_{k-1}|\,(a_{kk} - \mathbf{a}'\mathbf{A}_{k-1}^{-1}\mathbf{a})$$

Since by assumption $|\mathbf{A}_k| > 0$, and $|\mathbf{A}_{k-1}| > 0$, we obtain

$$a_{kk} - \mathbf{a}'\mathbf{A}_{k-1}^{-1}\mathbf{a} > 0$$

Then

$$\mathbf{x}' \left[\begin{array}{c|c} \mathbf{A}_{k-1} & \mathbf{0} \\ \hline \mathbf{0} & a_{kk} - \mathbf{a}'\mathbf{A}_{k-1}^{-1}\mathbf{a} \end{array}\right] \mathbf{x}$$

$$= [x_1\, x_2 \cdots x_{k-1}]\, \mathbf{A}_{k-1} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{k-1} \end{bmatrix} + (a_{kk} - \mathbf{a}'\mathbf{A}_{k-1}^{-1}\mathbf{a})x_k^2$$

Since $\mathbf{A}_{k-1}$ is assumed to be positive definite and $a_{kk} - \mathbf{a}'\mathbf{A}_{k-1}^{-1}\mathbf{a} > 0$, we conclude that

$$\mathbf{x}' \left[\begin{array}{c|c} \mathbf{A}_{k-1} & \mathbf{0} \\ \hline \mathbf{0} & a_{kk} - \mathbf{a}'\mathbf{A}_{k-1}^{-1}\mathbf{a} \end{array}\right] \mathbf{x} > 0 \qquad \text{for } \mathbf{x} \neq \mathbf{0}$$

Hence, $\mathbf{x}'\mathbf{A}\mathbf{x}$ is also positive definite. This proves the theorem.

We shall next present Sylvester's criterion for negative definiteness of quadratic forms.

Theorem 5-14 (Sylvester's criterion for negative definiteness of quadratic form) A necessary and sufficient condition in order that the quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is an $n \times n$ real symmetric matrix, be negative definite is that $\det \mathbf{A}$ be positive if n is even and negative if n is odd, and the successive principal minors of even order be positive and the successive principal minors of odd order be negative, or

$$a_{11} < 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0, \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} < 0, \quad \cdots$$

$$\det \mathbf{A} > 0 \quad \text{if} \quad n = \text{even}$$
$$\det \mathbf{A} < 0 \quad \text{if} \quad n = \text{odd}$$
$$(a_{ij} = a_{ji})$$

Proof The quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ is negative definite if $\mathbf{x}'(-\mathbf{A})\mathbf{x}$ is positive definite. A necessary and sufficient condition that $\mathbf{x}'(-\mathbf{A})\mathbf{x}$ is positive definite is given by

$$-a_{11} > 0, \quad \begin{vmatrix} -a_{11} & -a_{12} \\ -a_{21} & -a_{22} \end{vmatrix} > 0, \quad \begin{vmatrix} -a_{11} & -a_{12} & -a_{13} \\ -a_{21} & -a_{22} & -a_{23} \\ -a_{31} & -a_{32} & -a_{33} \end{vmatrix} > 0, \quad \cdots$$

$$\det(-\mathbf{A}) > 0$$

It follows that

$$a_{11} < 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0, \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} < 0, \quad \cdots$$

$$\det \mathbf{A} > 0 \quad \text{if} \quad n = \text{even}$$
$$\det \mathbf{A} < 0 \quad \text{if} \quad n = \text{odd}$$

Hence the theorem.

We shall next present Sylvester's criteria for semidefiniteness of quadratic forms.

Theorem 5-15 (Sylvester's criterion for positive semidefiniteness of quadratic form) A necessary and sufficient condition in order that the quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a real symmetric matrix, be positive semidefinite is that $\mathbf{A}$ be singular (det $\mathbf{A} = 0$) and all the principal minors be nonnegative, or

$$a_{ii} \geqslant 0, \quad \begin{vmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{vmatrix} \geqslant 0, \quad \begin{vmatrix} a_{ii} & a_{ij} & a_{ik} \\ a_{ji} & a_{jj} & a_{jk} \\ a_{ki} & a_{kj} & a_{kk} \end{vmatrix} \geqslant 0, \cdots$$

$$\det \mathbf{A} = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} = 0 \qquad (a_{ij} = a_{ji})$$

Theorem 5-16 (Sylvester's criterion for negative semidefiniteness of quadratic form) A necessary and sufficient condition in order that the quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a real symmetric matrix, be negative semidefinite is that $\mathbf{A}$ be singular (det $\mathbf{A} = 0$) and all the principal minors of even order be nonnegative and those of odd orders be nonpositive, or

$$a_{ii} \leqslant 0, \quad \begin{vmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{vmatrix} \geqslant 0, \quad \begin{vmatrix} a_{ii} & a_{ij} & a_{ik} \\ a_{ji} & a_{jj} & a_{jk} \\ a_{ki} & a_{kj} & a_{kk} \end{vmatrix} \leqslant 0, \cdots$$

$$\det \mathbf{A} = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} = 0 \qquad (a_{ij} = a_{ji})$$

Proofs of the last two theorems are similar to those of Theorems 5–13 and 5–14. We shall leave proofs to the reader.

As an example of application of Sylvester's criteria, consider the following quadratic form:

$$10x_1^2 + x_2^2 + 4x_3^2 + 4x_1x_2 + 6x_1x_3 + 2x_2x_3 = \mathbf{x}'\mathbf{A}\mathbf{x}$$

where

$$\mathbf{A} = \begin{bmatrix} 10 & 2 & 3 \\ 2 & 1 & 1 \\ 3 & 1 & 4 \end{bmatrix}$$

For this matrix $\mathbf{A}$, det $\mathbf{A} > 0$ and the successive principal minors are positive, or

$$10 > 0, \quad \begin{vmatrix} 10 & 2 \\ 2 & 1 \end{vmatrix} > 0, \quad \begin{vmatrix} 10 & 2 & 3 \\ 2 & 1 & 1 \\ 3 & 1 & 4 \end{vmatrix} > 0$$

Hence $\mathbf{x}'\mathbf{A}\mathbf{x}$ is positive definite.

As an another example, consider the following quadratic form:

$$x_1^2 - 3x_2^2 + x_3^2 + 4x_1x_2 + 8x_1x_3$$

The matrix $\mathbf{A}$ associated with this quadratic form is

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 4 \\ 2 & -3 & 0 \\ 4 & 0 & 1 \end{bmatrix}$$

Since

$$1 > 0, \quad \begin{vmatrix} 1 & 2 \\ 2 & -3 \end{vmatrix} < 0, \quad \begin{vmatrix} 1 & 2 & 4 \\ 2 & -3 & 0 \\ 4 & 0 & 1 \end{vmatrix} > 0$$

$\mathbf{x}'\mathbf{A}\mathbf{x}$ is neither positive definite nor negative definite. Since det $\mathbf{A} > 0$, $\mathbf{A}$ is nonsingular and cannot be semidefinite. Hence we conclude that $\mathbf{A}$ is indefinite.

We have presented Sylvester's criteria for quadratic forms. However, the same criteria can be derived for Hermitian forms. Theorems 5–13—5–16 hold true for Hermitian forms with trivial modifications that $\mathbf{x}'\mathbf{A}\mathbf{x}$ be changed to $\mathbf{x}^*\mathbf{A}\mathbf{x}$, $\mathbf{A}$ be changed from a real symmetric matrix to a Hermitian matrix, and $a_{ij} = a_{ji}$ be changed to $a_{ij} = \bar{a}_{ji}$.

EXERCISE PROBLEMS AND SOLUTIONS

Problem A-5-1 Obtain the minimal polynomials of the following matrices:

(1) $\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$

(2) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

Solution

(1) $(\lambda - 1)^2(\lambda - 2) = 0$

(2) $\lambda - 1 = 0$

Problem A-5-2 Prove that any two polynomials in an $n \times n$ matrix $\mathbf{A}$ commute, that is,

$$f(\mathbf{A})g(\mathbf{A}) = g(\mathbf{A})f(\mathbf{A})$$

where

$$f(\mathbf{A}) = \sum_{i=0}^{r} a_i \mathbf{A}^i \quad \text{and} \quad g(\mathbf{A}) = \sum_{j=0}^{s} b_j \mathbf{A}^j$$

Solution

$$f(\mathbf{A})g(\mathbf{A}) = \left(\sum_{i=0}^{r} a_i \mathbf{A}^i\right)\left(\sum_{j=0}^{s} b_j \mathbf{A}^j\right)$$

$$= \sum_{i=0}^{r}\sum_{j=0}^{s} a_i \mathbf{A}^i b_j \mathbf{A}^j = \sum_{i=0}^{r}\sum_{j=0}^{s} a_i b_j \mathbf{A}^{i+j}$$

$$= \left(\sum_{j=0}^{s} b_j \mathbf{A}^j\right)\left(\sum_{i=0}^{r} a_i \mathbf{A}^i\right) = g(\mathbf{A})f(\mathbf{A})$$

Problem A-5-3 Prove that the minimal polynomial $\phi_{\mathbf{A}}(\lambda)$ of an $n \times n$ matrix $\mathbf{A}$ and the minimal polynomial $\phi_{\mathbf{T}^{-1}\mathbf{A}\mathbf{T}}(\lambda)$ of $\mathbf{T}^{-1}\mathbf{A}\mathbf{T}$ are the same. ($\mathbf{T}$ is a nonsingular matrix.)

Solution Notice that

$$\phi_{\mathbf{A}}(\mathbf{T}^{-1}\mathbf{A}\mathbf{T}) = \mathbf{T}^{-1}[\phi_{\mathbf{A}}(\mathbf{A})]\,\mathbf{T} = \mathbf{0}$$

Hence $\phi_{\mathbf{A}}$ is divisible by $\phi_{\mathbf{T}^{-1}\mathbf{A}\mathbf{T}}$. Since

$$\mathbf{A} = \mathbf{T}(\mathbf{T}^{-1}\mathbf{A}\mathbf{T})\mathbf{T}^{-1}$$

we have

$$\phi_{\mathbf{T}^{-1}\mathbf{A}\mathbf{T}}(\mathbf{A})$$

$$= \phi_{\mathbf{T}^{-1}\mathbf{A}\mathbf{T}}[\mathbf{T}(\mathbf{T}^{-1}\mathbf{A}\mathbf{T})\mathbf{T}^{-1}]$$

$$= \mathbf{T}[\phi_{\mathbf{T}^{-1}\mathbf{A}\mathbf{T}}(\mathbf{T}^{-1}\mathbf{A}\mathbf{T})]\mathbf{T}^{-1} = \mathbf{0}$$

Hence $\phi_{\mathbf{T}^{-1}\mathbf{A}\mathbf{T}}$ is divisible by $\phi_{\mathbf{A}}$. Since the coefficients of the highest-degree term in both $\phi_{\mathbf{A}}$ and $\phi_{\mathbf{T}^{-1}\mathbf{A}\mathbf{T}}$ are unity, we conclude that

$$\phi_{\mathbf{T}^{-1}\mathbf{A}\mathbf{T}}(\lambda) = \phi_{\mathbf{A}}(\lambda)$$

Problem A-5-4 Suppose that $\lambda_1, \lambda_2, \ldots, \lambda_m$ are distinct roots of the minimal polynomial of an $n \times n$ matrix $\mathbf{A}$ and that the minimal polynomial of $\mathbf{A}$ is a product of distinct linear factors. By solving the following equation for $f(\mathbf{A})$:

$$\begin{vmatrix} 1 & 1 & \cdots & 1 & \mathbf{I} \\ \lambda_1 & \lambda_2 & \cdots & \lambda_m & \mathbf{A} \\ \lambda_1^2 & \lambda_2^2 & \cdots & \lambda_m^2 & \mathbf{A}^2 \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \lambda_1^{m-1} & \lambda_2^{m-1} & \cdots & \lambda_m^{m-1} & \mathbf{A}^{m-1} \\ f(\lambda_1) & f(\lambda_2) & \cdots & f(\lambda_m) & f(\mathbf{A}) \end{vmatrix} = 0$$

we can obtain

$$f(\mathbf{A}) = \sum_{k=1}^{m} f(\lambda_k) \frac{(\mathbf{A} - \lambda_1\mathbf{I}) \cdots (\mathbf{A} - \lambda_{k-1}\mathbf{I})(\mathbf{A} - \lambda_{k+1}\mathbf{I}) \cdots (\mathbf{A} - \lambda_m\mathbf{I})}{(\lambda_k - \lambda_1) \cdots (\lambda_k - \lambda_{k-1})(\lambda_k - \lambda_{k+1}) \cdots (\lambda_k - \lambda_m)}$$

Verify this for $m = 4$.

Solution By expanding the following matrix about the last row, we obtain

$$\Delta \equiv \begin{vmatrix} 1 & 1 & 1 & 1 & \mathbf{I} \\ \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 & \mathbf{A} \\ \lambda_1^2 & \lambda_2^2 & \lambda_3^2 & \lambda_4^2 & \mathbf{A}^2 \\ \lambda_1^3 & \lambda_2^3 & \lambda_3^3 & \lambda_4^3 & \mathbf{A}^3 \\ f(\lambda_1) & f(\lambda_2) & f(\lambda_3) & f(\lambda_4) & f(\mathbf{A}) \end{vmatrix}$$

$$= f(\mathbf{A}) \begin{vmatrix} 1 & 1 & 1 & 1 \\ \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 \\ \lambda_1^2 & \lambda_2^2 & \lambda_3^2 & \lambda_4^2 \\ \lambda_1^3 & \lambda_2^3 & \lambda_3^3 & \lambda_4^3 \end{vmatrix} - f(\lambda_4) \begin{vmatrix} 1 & 1 & 1 & \mathbf{I} \\ \lambda_1 & \lambda_2 & \lambda_3 & \mathbf{A} \\ \lambda_1^2 & \lambda_2^2 & \lambda_3^2 & \mathbf{A}^2 \\ \lambda_1^3 & \lambda_2^3 & \lambda_3^3 & \mathbf{A}^3 \end{vmatrix}$$

$$+ f(\lambda_3) \begin{vmatrix} 1 & 1 & 1 & \mathbf{I} \\ \lambda_1 & \lambda_2 & \lambda_4 & \mathbf{A} \\ \lambda_1^2 & \lambda_2^2 & \lambda_4^2 & \mathbf{A}^2 \\ \lambda_1^3 & \lambda_2^3 & \lambda_4^3 & \mathbf{A}^3 \end{vmatrix} - f(\lambda_2) \begin{vmatrix} 1 & 1 & 1 & \mathbf{I} \\ \lambda_1 & \lambda_3 & \lambda_4 & \mathbf{A} \\ \lambda_1^2 & \lambda_3^2 & \lambda_4^2 & \mathbf{A}^2 \\ \lambda_1^3 & \lambda_3^3 & \lambda_4^3 & \mathbf{A}^3 \end{vmatrix}$$

$$+ f(\lambda_1) \begin{vmatrix} 1 & 1 & 1 & \mathbf{I} \\ \lambda_2 & \lambda_3 & \lambda_4 & \mathbf{A} \\ \lambda_2^2 & \lambda_3^2 & \lambda_4^2 & \mathbf{A}^2 \\ \lambda_2^3 & \lambda_3^3 & \lambda_4^3 & \mathbf{A}^3 \end{vmatrix}$$

Since Vandermonde's determinant has the property that

$$\begin{vmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_m \\ \lambda_1^2 & \lambda_2^2 & \cdots & \lambda_m^2 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \lambda_1^{m-1} & \lambda_2^{m-1} & \cdots & \lambda_m^{m-1} \end{vmatrix} = \prod_{1 \le i < j \le m} (\lambda_j - \lambda_i)$$

we obtain

$$\begin{vmatrix} 1 & 1 & 1 & 1 \\ \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 \\ \lambda_1^2 & \lambda_2^2 & \lambda_3^2 & \lambda_4^2 \\ \lambda_1^3 & \lambda_2^3 & \lambda_3^3 & \lambda_4^3 \end{vmatrix} = (\lambda_4 - \lambda_3)(\lambda_4 - \lambda_2)(\lambda_4 - \lambda_1)(\lambda_3 - \lambda_2)(\lambda_3 - \lambda_1)(\lambda_2 - \lambda_1)$$

and

$$\begin{vmatrix} 1 & 1 & 1 & \mathbf{I} \\ \lambda_i & \lambda_j & \lambda_k & \mathbf{A} \\ \lambda_i^2 & \lambda_j^2 & \lambda_k^2 & \mathbf{A}^2 \\ \lambda_i^3 & \lambda_j^3 & \lambda_k^3 & \mathbf{A}^3 \end{vmatrix} = (\mathbf{A} - \lambda_k\mathbf{I})(\mathbf{A} - \lambda_j\mathbf{I})(\mathbf{A} - \lambda_i\mathbf{I})(\lambda_k - \lambda_j)(\lambda_k - \lambda_i)(\lambda_j - \lambda_i)$$

Therefore, we obtain

$$\begin{aligned} \Delta = {} & f(\mathbf{A})\,[(\lambda_4 - \lambda_3)(\lambda_4 - \lambda_2)(\lambda_4 - \lambda_1)(\lambda_3 - \lambda_2)(\lambda_3 - \lambda_1)(\lambda_2 - \lambda_1)] \\ & - f(\lambda_4)\,[(\mathbf{A} - \lambda_3\mathbf{I})(\mathbf{A} - \lambda_2\mathbf{I})(\mathbf{A} - \lambda_1\mathbf{I})(\lambda_3 - \lambda_2)(\lambda_3 - \lambda_1)(\lambda_2 - \lambda_1)] \\ & + f(\lambda_3)\,[(\mathbf{A} - \lambda_4\mathbf{I})(\mathbf{A} - \lambda_2\mathbf{I})(\mathbf{A} - \lambda_1\mathbf{I})(\lambda_4 - \lambda_2)(\lambda_4 - \lambda_1)(\lambda_2 - \lambda_1)] \\ & - f(\lambda_2)\,[(\mathbf{A} - \lambda_4\mathbf{I})(\mathbf{A} - \lambda_3\mathbf{I})(\mathbf{A} - \lambda_1\mathbf{I})(\lambda_4 - \lambda_3)(\lambda_4 - \lambda_1)(\lambda_3 - \lambda_1)] \\ & + f(\lambda_1)\,[(\mathbf{A} - \lambda_4\mathbf{I})(\mathbf{A} - \lambda_3\mathbf{I})(\mathbf{A} - \lambda_2\mathbf{I})(\lambda_4 - \lambda_3)(\lambda_4 - \lambda_2)(\lambda_3 - \lambda_2)] \\ = {} & 0 \end{aligned}$$

Solving for $f(\mathbf{A})$, we get

$$\begin{aligned} f(\mathbf{A}) = {} & f(\lambda_1)\frac{(\mathbf{A} - \lambda_2\mathbf{I})(\mathbf{A} - \lambda_3\mathbf{I})(\mathbf{A} - \lambda_4\mathbf{I})}{(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3)(\lambda_1 - \lambda_4)} + f(\lambda_2)\frac{(\mathbf{A} - \lambda_1\mathbf{I})(\mathbf{A} - \lambda_3\mathbf{I})(\mathbf{A} - \lambda_4\mathbf{I})}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3)(\lambda_2 - \lambda_4)} \\ & + f(\lambda_3)\frac{(\mathbf{A} - \lambda_1\mathbf{I})(\mathbf{A} - \lambda_2\mathbf{I})(\mathbf{A} - \lambda_4\mathbf{I})}{(\lambda_3 - \lambda_1)(\lambda_3 - \lambda_2)(\lambda_3 - \lambda_4)} + f(\lambda_4)\frac{(\mathbf{A} - \lambda_1\mathbf{I})(\mathbf{A} - \lambda_2\mathbf{I})(\mathbf{A} - \lambda_3\mathbf{I})}{(\lambda_4 - \lambda_1)(\lambda_4 - \lambda_2)(\lambda_4 - \lambda_3)} \\ = {} & \sum_{k=1}^{m} f(\lambda_k)\frac{(\mathbf{A} - \lambda_1\mathbf{I}) \cdots (\mathbf{A} - \lambda_{k-1}\mathbf{I})(\mathbf{A} - \lambda_{k+1}\mathbf{I}) \cdots (\mathbf{A} - \lambda_m\mathbf{I})}{(\lambda_k - \lambda_1) \cdots (\lambda_k - \lambda_{k-1})(\lambda_k - \lambda_{k+1}) \cdots (\lambda_k - \lambda_m)} \end{aligned}$$

where $m = 4$.

Problem A-5-5 Prove that, for an $n \times n$ constant matrix $\mathbf{A}$, the series defining the exponential of $\mathbf{A}t$, namely,

$$e^{\mathbf{A}t} = \mathbf{I} + \mathbf{A}t + \frac{1}{2!}\mathbf{A}^2t^2 + \cdots + \frac{1}{m!}\mathbf{A}^m t^m + \cdots$$

is convergent for all $\mathbf{A}$ and for all finite t.

Solution Let

$$\mathbf{A}t = (a_{ij}t) \qquad \text{and} \qquad \max_{i,j} |a_{ij}t| = M$$

Put

$$\mathbf{A}^m t^m = (a_{ij}t)^{(m)}$$

then we can prove that

$$|(a_{ij}t)^{(m)}| \leqslant n^{m-1}M^m \qquad (1 \leqslant i \leqslant n\ ;\ 1 \leqslant j \leqslant n) \tag{5-23}$$

To prove inequality (5–23) let us use induction. Clearly this is true for $m = 1$. If inequality (5–23) is true for $m - 1$, then

$$|(a_{ij}t)^{(m-1)}| \leqslant n^{m-2}M^{m-1}$$

Hence

$$\begin{aligned} |(a_{ij}t)^{(m)}| &= \left|\sum_{k=1}^{n}(a_{ik}t)^{(m-1)}a_{kj}t\right| \\ &\leqslant \sum_{k=1}^{n}|(a_{ik}t)^{(m-1)}||a_{kj}t| \\ &\leqslant n \cdot n^{m-2}M^{m-1}M = n^{m-1}M^m \end{aligned}$$

Therefore inequality (5–23) holds for m. Since for all finite M

$$\frac{1}{n}e^{nM} = \sum_{m=0}^{\infty}\frac{1}{m!}n^{m-1}M^m$$

converges, we see that the series

$$\sum_{m=0}^{\infty}\frac{1}{m!}(a_{ij}t)^{(m)}$$

also converges for all finite t. Therefore, the series

$$\mathbf{I} + \mathbf{A}t + \frac{1}{2!}\mathbf{A}^2t^2 + \cdots + \frac{1}{m!}\mathbf{A}^m t^m + \cdots$$

converges for all constant matrix $\mathbf{A}$ and for all finite t.

Problem A-5-6 Show that

(1) $(\mathbf{A}^*)^k = (\mathbf{A}^k)^* \qquad (k = 1, 2, \ldots)$

(2) $(e^{\mathbf{A}t})^* = e^{\mathbf{A}^*t}$

where $\mathbf{A}$ is an $n \times n$ constant matrix and t is a scalar.

Solution For (1), note that $(\mathbf{A}^*)^k$ can be written as

$$(\mathbf{A}^*)^k = \mathbf{A}_1^*\mathbf{A}_2^*\mathbf{A}_3^* \cdots \mathbf{A}_{k-1}^*\mathbf{A}_k^*$$

where the subscript identifies the mth matrix $\mathbf{A}$. Since the conjugate transpose of the product of matrices is equal to the product of their conjugate transposes in the reverse order; that is, $(\mathbf{ABC})^* = \mathbf{C}^*\mathbf{B}^*\mathbf{A}^*$

$$\begin{aligned} (\mathbf{A}^*)^k &= (\mathbf{A}_k\mathbf{A}_{k-1} \cdots \mathbf{A}_3\mathbf{A}_2\mathbf{A}_1)^* \\ &= (\mathbf{A}^k)^* \end{aligned}$$

Thus we see that $(\mathbf{A}^*)^k = (\mathbf{A}^k)^*$.

For (2), $e^{\mathbf{A}t}$ can be written as

$$e^{\mathbf{A}t} = \mathbf{I} + \sum_{k=1}^{\infty}\frac{(\mathbf{A}t)^k}{k!}$$

Then, taking the conjugate transpose of both sides of this last equation

$$\begin{aligned} (e^{\mathbf{A}t})^* &= \left(\mathbf{I} + \sum_{k=1}^{\infty}\frac{(\mathbf{A}t)^k}{k!}\right)^* \\ &= \mathbf{I} + \sum_{k=1}^{\infty}\frac{(\mathbf{A}^k)^*t^k}{k!} \end{aligned}$$

but part (1) showed that

$$(\mathbf{A}^k)^* = (\mathbf{A}^*)^k$$

Hence,

$$(e^{\mathbf{A}t})^* = \mathbf{I} + \sum_{k=1}^{\infty} \frac{(\mathbf{A}^*)^k t^k}{k!} = e^{\mathbf{A}^* t}$$

Problem A-5-7 Compute $e^{\mathbf{A}t}$ where

$$\mathbf{A} = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}$$

Solution

$$\begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix} = \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

Since the matrices on the right-hand side of the preceding equation commute,

$$\exp \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix} t = \left\{ \exp \begin{bmatrix} \lambda t & 0 & 0 \\ 0 & \lambda t & 0 \\ 0 & 0 & \lambda t \end{bmatrix} \right\} \left\{ \exp \begin{bmatrix} 0 & t & 0 \\ 0 & 0 & t \\ 0 & 0 & 0 \end{bmatrix} \right\}$$

Since

$$\exp \begin{bmatrix} 0 & t & 0 \\ 0 & 0 & t \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & t & \frac{1}{2}t^2 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix}$$

we have

$$\exp \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix} t = e^{\lambda t} \begin{bmatrix} 1 & t & \frac{1}{2}t^2 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix}$$

Problem A-5-8 Compute $e^{\mathbf{A}t}$ where

$$\mathbf{A} = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{bmatrix}$$

Solution Let

$$\mathbf{B}t = \begin{bmatrix} \lambda t & & & 0 \\ & \lambda t & & \\ & & \cdot & \\ & & & \cdot \\ 0 & & & \lambda t \end{bmatrix}, \qquad \mathbf{C}t = \begin{bmatrix} 0 & t & 0 & \cdots & 0 \\ 0 & 0 & t & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

Then we obtain

$$\begin{aligned} e^{\mathbf{A}t} &= e^{\mathbf{B}t} e^{\mathbf{C}t} \\ &= e^{\mathbf{B}t} \left[\mathbf{I} + \mathbf{C}t + \frac{1}{2!}(\mathbf{C}t)^2 + \cdots + \frac{1}{(n-1)!}(\mathbf{C}t)^{n-1} \right] \\ &= e^{\lambda t} \begin{bmatrix} 1 & t & \frac{t^2}{2!} & \cdots & \frac{t^{n-1}}{(n-1)!} \\ 0 & 1 & t & \cdots & \frac{t^{n-2}}{(n-2)!} \\ \cdot & \cdot & \cdot & & \vdots \\ \cdot & \cdot & \cdot & & \vdots \\ \cdot & \cdot & \cdot & & \vdots \\ 0 & 0 & 0 & \cdots & t \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \end{aligned}$$

Problem A-5-9 The derivative of $|\mathbf{A}(t)|$ can be given by

$$\frac{d}{dt}|\mathbf{A}(t)| = \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_{ij} \frac{d}{dt} a_{ij}(t)$$

Since

$$\mathbf{A}^{-1}(t) = \frac{1}{|\mathbf{A}(t)|}(\Delta_{ji})$$

we have

$$\frac{1}{|\mathbf{A}(t)|} \frac{d|\mathbf{A}(t)|}{dt} = \operatorname{tr}\left(\mathbf{A}^{-1}(t) \frac{d}{dt} \mathbf{A}(t) \right)$$

By use of this relationship show that, for a constant matrix A,

$$|e^{\mathbf{A}}| = e^{\operatorname{tr}\mathbf{A}}$$

(Hence $|e^{\mathbf{A}}| = 1$ implies that $\operatorname{tr}\mathbf{A} = 0$.)

Solution Let $\mathbf{A}(t) = e^{\mathbf{A}t}$. Then

$$\frac{d}{dt}\mathbf{A}(t) = \frac{d}{dt} e^{\mathbf{A}t} = \mathbf{A}e^{\mathbf{A}t} = e^{\mathbf{A}t}\mathbf{A}$$

and

$$\frac{1}{|e^{\mathbf{A}t}|} \frac{d|e^{\mathbf{A}t}|}{dt} = \operatorname{tr}\mathbf{A}$$

Hence

$$|e^{\mathbf{A}t}| = Ce^{(\operatorname{tr}\mathbf{A})t}$$

By substituting $t = 0$ in this last equation, we find $C = 1$. By letting $t = 1$,

$$|e^{\mathbf{A}}| = e^{\operatorname{tr}\mathbf{A}}$$

Problem A-5-10 For a real symmetric matrix A with nonzero eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ show that

$$\mathbf{A}^{-1} = \mathbf{T} \begin{bmatrix} \lambda_1^{-1} & & & 0 \\ & \lambda_2^{-1} & & \\ & & \cdot & \\ & & & \cdot \\ 0 & & & \lambda_n^{-1} \end{bmatrix} \mathbf{T}' \qquad (\mathbf{T} = \text{orthogonal matrix})$$

Solution By Theorem 5–7, we may write

$$\mathbf{T}'\mathbf{A}\mathbf{T} = \mathbf{D} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

Then,

$$(\mathbf{T}'\mathbf{A}\mathbf{T})^{-1} = \mathbf{T}^{-1}\mathbf{A}^{-1}(\mathbf{T}')^{-1} = \mathbf{D}^{-1}$$

and we obtain

$$\mathbf{T}\mathbf{T}^{-1}\mathbf{A}^{-1}(\mathbf{T}')^{-1}\mathbf{T}' = \mathbf{A}^{-1} = \mathbf{T}\mathbf{D}^{-1}\mathbf{T}'$$

Problem A-5-11 For a positive definite matrix **A** there exists an orthogonal matrix **T** such that

$$\mathbf{T}'\mathbf{A}\mathbf{T} = \mathbf{D} = \text{diag}\,(d_1, d_2, \ldots, d_n)$$

where the d_k are all positive. The matrix **T** defines an orthogonal transformation of coordinates reducing the quadratic form, $\mathbf{x}'\mathbf{A}\mathbf{x}$ = constant, to its axes. Show that if we put

$$\mathbf{T}'\mathbf{A}^{-1}\mathbf{T} = \mathbf{\Gamma}$$

then

$$\mathbf{D}\mathbf{\Gamma} = \text{diag}\,(1, 1, \ldots, 1)$$

(Note that

$$\mathbf{\Gamma} = \mathbf{D}^{-1} = \text{diag}\left(\frac{1}{d_1}, \frac{1}{d_2}, \ldots, \frac{1}{d_n}\right)$$

is also positive definite.)

Solution

$$\mathbf{D}\mathbf{\Gamma} = \mathbf{T}'\mathbf{A}\mathbf{T}\mathbf{T}'\mathbf{A}^{-1}\mathbf{T} = \mathbf{T}^{-1}\mathbf{A}\mathbf{A}^{-1}\mathbf{T} = \mathbf{I} = \text{diag}\,(1, 1, \ldots, 1)$$

Problem A-5-12 Show that if a scalar quantity f is defined by

$$f = \mathbf{x}'\mathbf{A}\mathbf{x}$$

where **A** is an $n \times n$ real symmetric matrix, then

$$\text{grad}\, f = 2\mathbf{A}\mathbf{x}$$

Solution If we write f in the following form:

$$f = [x_1\ x_2\ \cdots\ x_n]\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$

$$= [x_1\ x_2\ \cdots\ x_n]\begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \cdot \\ \cdot \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n \end{bmatrix}$$

$$= (a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n)\,x_1 + (a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n)\,x_2 + \cdots + (a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n)\,x_n$$

where $a_{ij} = a_{ji}$, we find

$$\frac{\partial f}{\partial x_i} = 2(a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n)$$

$$(i = 1, 2, \ldots, n)$$

Hence

$$\text{grad}\, f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \cdot \\ \cdot \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = 2\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = 2\mathbf{A}\mathbf{x}$$

Problem A-5-13 The kinetic energy of a system is given by

$$T = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\dot{q}_i\dot{q}_j$$

where the a_{ij} are constants or functions of the q_i and $a_{ij} = a_{ji}$. Show that if p_i is defined by

$$p_i = \frac{\partial T}{\partial \dot{q}_i}$$

then

$$\dot{q}_i = \frac{\partial T}{\partial p_i}$$

Solution The kinetic energy T can be written as

$$T = \tfrac{1}{2}\dot{\mathbf{q}}'\mathbf{A}\dot{\mathbf{q}}$$

where

$$\dot{\mathbf{q}} = \begin{bmatrix} \dot{q}_1 \\ \dot{q}_2 \\ \cdot \\ \cdot \\ \dot{q}_n \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \qquad (a_{ij} = a_{ji})$$

Since $T = \frac{1}{2}\dot{\mathbf{q}}'\mathbf{A}\dot{\mathbf{q}}$ and p_i is defined by

$$p_i = \frac{\partial T}{\partial \dot{q}_i}$$

we obtain (see Prob. A-5-12)

$$\mathbf{p} = (\text{grad}\, T)_{\dot{q}} = \mathbf{A}\dot{\mathbf{q}}$$

Hence

$$\dot{\mathbf{q}} = \mathbf{A}^{-1}\mathbf{p}$$

(Since the kinetic energy T is a positive definite quantity, **A** is nonsingular. Hence $\mathbf{A}^{-1}$ exists.) We thus obtain

$$T = \tfrac{1}{2}\dot{\mathbf{q}}'\mathbf{A}\dot{\mathbf{q}} = \tfrac{1}{2}(\mathbf{A}^{-1}\mathbf{p})'\mathbf{A}(\mathbf{A}^{-1}\mathbf{p})$$
$$= \tfrac{1}{2}(\mathbf{A}^{-1}\mathbf{p})'\mathbf{p} = \tfrac{1}{2}\mathbf{p}'\mathbf{A}^{-1}\mathbf{p}$$

Therefore,

$$(\text{grad } T)_p = \mathbf{A}^{-1}\mathbf{p}$$

Hence

$$\dot{\mathbf{q}} = (\text{grad } T)_p$$

or

$$\dot{q}_i = \frac{\partial T}{\partial p_i}$$

Problem A-5-14 Show that if

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{A}_{n-1} & \mathbf{a} \\ \hline \mathbf{a}' & a_{nn} \end{array}\right]$$

then

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline \mathbf{a}'\mathbf{A}_{n-1}^{-1} & 1 \end{array}\right]\left[\begin{array}{c|c} \mathbf{A}_{n-1} & \mathbf{0} \\ \hline \mathbf{0} & a_{nn} - \mathbf{a}'\mathbf{A}_{n-1}^{-1}\mathbf{a} \end{array}\right]\left[\begin{array}{c|c} \mathbf{I} & \mathbf{A}_{n-1}^{-1}\mathbf{a} \\ \hline \mathbf{0} & 1 \end{array}\right]$$

Solution

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline \mathbf{a}'\mathbf{A}_{n-1}^{-1} & 1 \end{array}\right]\left[\begin{array}{c|c} \mathbf{A}_{n-1} & \mathbf{a} \\ \hline \mathbf{0} & a_{nn} - \mathbf{a}'\mathbf{A}_{n-1}^{-1}\mathbf{a} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A}_{n-1} & \mathbf{a} \\ \hline \mathbf{a}' & a_{nn} \end{array}\right]$$

Problem A-5-15 Show that the following quadratic form:

$$-x_1^2 - 10x_2^2 - 4x_3^2 + 6x_1x_2 + 2x_2x_3$$

is negative definite.

Solution The given quadratic form can be written as

$$\mathbf{x}'\mathbf{A}\mathbf{x} = [x_1\ x_2\ x_3]\begin{bmatrix} -1 & 3 & 0 \\ 3 & -10 & 1 \\ 0 & 1 & -4 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

The successive principal minors and the determinant are

$$-1 < 0, \quad \begin{vmatrix} -1 & 3 \\ 3 & -10 \end{vmatrix} > 0, \quad \begin{vmatrix} -1 & 3 & 0 \\ 3 & -10 & 1 \\ 0 & 1 & -4 \end{vmatrix} < 0$$

By Theorem 5–14, $\mathbf{x}'\mathbf{A}\mathbf{x}$ is negative definite.

PROBLEMS

Problem B-5-1 If the minimal polynomial of an $n \times n$ matrix $\mathbf{A}$ where

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{A}_1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{A}_2 \end{array}\right]$$

is defined by $\phi_{\mathbf{A}}(\lambda)$, then show that $\phi_{\mathbf{A}}(\lambda)$ is the minimum common factor of $\phi_{\mathbf{A}_1}(\lambda)$ and $\phi_{\mathbf{A}_2}(\lambda)$, where $\phi_{\mathbf{A}_1}(\lambda)$ and $\phi_{\mathbf{A}_2}(\lambda)$ are the minimal polynomial of $\mathbf{A}_1$ and that of $\mathbf{A}_2$, respectively.

Problem B-5-2 Show that if $\|\mathbf{A}\| < 1$, then

$$\sum_{i=0}^{\infty} \mathbf{A}^i = (\mathbf{I} - \mathbf{A})^{-1}$$

Problem B-5-3 Show that if

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ a & 0 & 0 & 0 & 0 \end{bmatrix}$$

then

$$\mathbf{A}^2 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ a & 0 & 0 & 0 & 0 \\ 0 & a & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{A}^3 = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ a & 0 & 0 & 0 & 0 \\ 0 & a & 0 & 0 & 0 \\ 0 & 0 & a & 0 & 0 \end{bmatrix}$$

$$\mathbf{A}^5 = \begin{bmatrix} a & & & & 0 \\ & a & & & \\ & & a & & \\ & & & a & \\ 0 & & & & a \end{bmatrix}, \quad \mathbf{A}^6 = \begin{bmatrix} 0 & a & 0 & 0 & 0 \\ 0 & 0 & a & 0 & 0 \\ 0 & 0 & 0 & a & 0 \\ 0 & 0 & 0 & 0 & a \\ a^2 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{A}^{10} = \begin{bmatrix} a^2 & & & & 0 \\ & a^2 & & & \\ & & a^2 & & \\ & & & a^2 & \\ 0 & & & & a^2 \end{bmatrix}, \quad \mathbf{A}^{15} = \begin{bmatrix} a^3 & & & & 0 \\ & a^3 & & & \\ & & a^3 & & \\ & & & a^3 & \\ 0 & & & & a^3 \end{bmatrix}$$

Notice that similar results hold for any $n \times n$ matrix $\mathbf{A}$ of the following form:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ a & 0 & 0 & \cdots & 0 \end{bmatrix}$$

Problem B-5-4 For a skew symmetric matrix $\mathbf{A}$, show that

$$(e^{\mathbf{A}t})' = e^{-\mathbf{A}t} \quad \text{or} \quad (e^{\mathbf{A}t})'(e^{\mathbf{A}t}) = \mathbf{I}$$

Problem B-5-5 Consider a square matrix $\mathbf{A}$ defined by

$$\mathbf{A} = \text{diag}\,(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_m)$$

where the $\mathbf{A}_i$ are square matrices of orders r_i $(i = 1, 2, \ldots, m)$. $\mathbf{A}^2$ is then given by

$$\mathbf{A}^2 = \text{diag}\,(\mathbf{A}_1^2, \mathbf{A}_2^2, \ldots, \mathbf{A}_m^2)$$

and similarly

$$\mathbf{A}^k = \text{diag}\,(\mathbf{A}_1^k, \mathbf{A}_2^k, \ldots, \mathbf{A}_m^k) \qquad (k = 1, 2, \ldots)$$

Hence we can obtain

$$f(\mathbf{A}) = \text{diag}\,(f(\mathbf{A}_1), f(\mathbf{A}_2), \ldots, f(\mathbf{A}_m))$$

where $f(\mathbf{A})$ is any polynomial in $\mathbf{A}$. Notice that if $f(\mathbf{A})$ is an infinite series, it converges to a limit if and only if all the $f(\mathbf{A}_i)$ converge to respective limits.

By use of the foregoing relationship obtain $\sin \mathbf{A}$ and $\cos \mathbf{A}$ where

$$\mathbf{A} = \begin{bmatrix} a_{11} & & & 0 \\ & a_{22} & & \\ & & \ddots & \\ 0 & & & a_{nn} \end{bmatrix}$$

Verify that

$$\sin^2 \mathbf{A} + \cos^2 \mathbf{A} = \mathbf{I}$$

Problem B-5-6 Prove that if the eigenvalues of an $n \times n$ matrix $\mathbf{A}$ are $\lambda_1, \lambda_2, \ldots, \lambda_n$, then the eigenvalues of $\mathbf{A}^k$ are $\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k$, where k is an integer.

Problem B-5-7 An $n \times n$ matrix $\mathbf{A}$ has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ which are not necessarily distinct. Show that there exists a matrix $\mathbf{T}$ such that

$$\mathbf{A} = \mathbf{T} \begin{bmatrix} \lambda_1 & & & * \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix} \mathbf{T}^{-1}$$

where $*$ indicates that the elements above the diagonal are not necessarily zero. By use of this show also that

$$\sum_{i=1}^{n} \lambda_i^k = \text{tr}\,(\mathbf{A}^k)$$

Problem B-5-8 Show that for a symmetric matrix $\mathbf{A} = (a_{ij})$

$$\frac{\partial^2}{\partial x_i \, \partial x_j} \mathbf{x}'\mathbf{A}\mathbf{x} = 2a_{ij}$$

Problem B-5-9 Suppose that a nonsingular matrix $\mathbf{P}$ transforms a real symmetric matrix $\mathbf{A}$ such that $\mathbf{P}'\mathbf{A}\mathbf{P} = \mathbf{I}$. Show that

$$|\mathbf{P}|^2 = \frac{1}{\prod_{i=1}^{n} \lambda_i} > 0$$

where the λ_i are the eigenvalues of $\mathbf{A}$.

Problem B-5-10 Prove that a Hermitian form $\mathbf{x}^*\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is an $n \times n$ Hermitian matrix having n distinct nonzero eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and $\mathbf{x}$ is a complex n vector, can be reduced to the diagonal form

$$\lambda_1 \bar{y}_1 y_1 + \lambda_2 \bar{y}_2 y_2 + \cdots + \lambda_n \bar{y}_n y_n$$

by a suitable unitary transformation $\mathbf{x} = \mathbf{T}\mathbf{y}$.

Problem B-5-11 Find a suitable transformation matrix $\mathbf{S}$ such that $\mathbf{x} = \mathbf{S}\mathbf{y}$ will simultaneously convert the following two quadratic forms:

$$Q_1 = \mathbf{x}'\mathbf{A}\mathbf{x} = 8.2x_1^2 + 4.8x_1x_2 + 6.8x_2^2 + 3x_3^2$$
$$Q_2 = \mathbf{x}'\mathbf{B}\mathbf{x} = x_1^2 + 2x_2^2 + 3x_3^2 + 2x_1x_2 + 2x_1x_3 + 2x_2x_3$$

into

$$Q_1 = y_1^2 + y_2^2 + y_3^2$$
$$Q_2 = ay_1^2 + by_2^2 + cy_3^2$$

where a, b, and c are constants. (Notice that the eigenvalues of $\mathbf{A}$ are 3, 5, 10, and, therefore, Q_1 is positive definite.)

Problem B-5-12 Prove that for a real symmetric matrix $\mathbf{A}$, the following statements are equivalent:

(1) $\mathbf{A}$ is positive definite.
(2) The determinant of $\mathbf{A}$ is positive and the successive principal minors of the determinant of $\mathbf{A}$ are positive.
(3) All eigenvalues of $\mathbf{A}$ are positive.
(4) There exists a nonsingular matrix $\mathbf{B}$ such that

$$\mathbf{B}'\mathbf{B} = \mathbf{A}$$

Problem B-5-13 Suppose that the elements a_{ij} of an $n \times n$ matrix $\mathbf{A}$ are non-negative. Prove that a necessary and sufficient condition that the absolute values of the eigenvalues of $\mathbf{A}$ are all smaller than unity is given by

$$\begin{vmatrix} 1 - a_{11} & -a_{12} & \cdots & -a_{1k} \\ -a_{21} & 1 - a_{22} & \cdots & -a_{2k} \\ \vdots & \vdots & & \vdots \\ -a_{k1} & -a_{k2} & \cdots & 1 - a_{kk} \end{vmatrix} > 0$$

$$(k = 1, 2, \ldots, n)$$

6

Obtaining the Solution of the State Space Equation

6-1 INTRODUCTION

State space representations of linear dynamic systems are very frequently in the form of linear vector matrix differential equations. The objective of this chapter is to present standard techniques for solving such linear vector matrix differential equations.[1] For numerical solutions by the use of digital computers, such linear vector matrix differential equations must be discretized. When discretized we will have vector matrix difference equations. Formal solutions of linear vector matrix difference equations are quite analogous to those of linear vector matrix differential equations. We, therefore, include discussions of formal solutions of linear vector matrix difference equations in this chapter.

The following sections of this chapter may be outlined as follows: Section 6–2 presents properties of the solutions of linear vector matrix differential equations. Formal solutions of such equations are derived in this section. Section 6–3 gives techniques for obtaining solutions of linear time-invariant vector matrix differential equations. It is shown that the solution will include the term $e^{\mathbf{A}t}$. In Section 6–4 we shall discuss methods for computing $e^{\mathbf{A}t}$ for a constant matrix $\mathbf{A}$. Section 6–5 presents transfer matrices and two methods for computing $(s\mathbf{I} - \mathbf{A})^{-1}$. Section 6–6 gives analyses of impulse responses of dynamic systems. Finally, in Section 6–7 we present formal solutions of linear vector matrix difference equations and discuss discretization of linear vector matrix differential equations.

[1]For detailed discussions of vector matrix differential equations see, for example (*6–1*, *6–2*). For applications of theory of differential equations to control systems see also (*6–3*). Figures in parentheses refer to books listed at the end of this chapter.

6-2 LINEAR VECTOR MATRIX DIFFERENTIAL EQUATIONS

In this section we shall obtain formal solutions of state space equations which are in the following form of vector matrix differential equation:

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u} \tag{6–1}$$

Here, we assume that $\mathbf{x}$ is an n vector; $\mathbf{u}$, an r vector; $\mathbf{A}(t)$, an $n \times n$ matrix; and $\mathbf{B}(t)$, an $n \times r$ matrix. We further assume that $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are absolutely integrable as functions of t in the interval $0 \leqslant t - t_0 \leqslant T$ or $t_0 \leqslant t \leqslant t_0 + T \equiv t_1$. Then, it can be proved that, for given t_0 and $\mathbf{x}(t_0)$, there exists a unique solution of Eq. (6–1) over the time interval $t_0 \leqslant t \leqslant t_1$.

In this book we shall frequently treat the case where the elements of $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are continuous functions of time. In such a case clearly $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are absolutely integrable as functions of time.

Basis functions and Wronskian Consider the linear homogeneous vector matrix differential equation

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} \tag{6–2}$$

where $\mathbf{x}$ is an n vector and $\mathbf{A}(t)$ is an $n \times n$ matrix whose elements are assumed to be absolutely integrable as functions of t in the interval $t_0 \leqslant t \leqslant t_1$. Assume that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are linearly independent solutions of Eq. (6–2). Any other solutions can be expressed as linear combinations of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. (Once we specify the initial condition the solution is uniquely determined.) We shall see later that a necessary and sufficient condition that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are linearly independent solutions of Eq. (6–2) is that the determinant of the matrix consisting of column vectors $\mathbf{x}_i$ is nonzero, or the following determinant is nonzero:

$$W = \begin{vmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{vmatrix} \neq 0$$

This determinant is called *Wronskian*. Wronskian plays an important role in theory of ordinary differential equations.

Entire solutions of Eq. (6–2) form an n-dimensional vector space. Any set of n linearly independent solutions constitutes a basis of the n-dimensional vector space. For a homogeneous scalar differential equation, any set of n linearly independent solutions is called a *set of basis functions*.

We shall next present several theorems on the solutions of linear differential equations.

Theorem 6-1 If n differentiable functions $x_1(t), x_2(t), \ldots, x_n(t)$ are linearly dependent for the time interval $t_0 \leqslant t \leqslant t_1$, then the Wronskian $W(x_1, x_2, \ldots, x_n)$ is zero for $t_0 \leqslant t \leqslant t_1$.

Proof The Wronskian $W(x_1, x_2, \ldots, x_n)$ is given by

$$W(x_1, x_2, \ldots, x_n) = \begin{vmatrix} x_1 & x_2 & \cdots & x_n \\ \dot{x}_1 & \dot{x}_2 & \cdots & \dot{x}_n \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_1^{(n-1)} & x_2^{(n-1)} & \cdots & x_n^{(n-1)} \end{vmatrix}$$

Since $x_1, x_2, \ldots, x_n$ are linearly dependent in the time interval $t_0 \leqslant t \leqslant t_1$, we have

$$c_1x_1(t) + c_2x_2(t) + \cdots + c_nx_n(t) = 0 \qquad (6\text{–}3)$$

where not all of $c_1, c_2, \ldots, c_n$ are zero. Differentiating successively Eq. (6–3) with respect to t, we have

$$\begin{aligned} c_1\dot{x}_1(t) + c_2\dot{x}_2(t) + \cdots + c_n\dot{x}_n(t) &= 0 \\ c_1\ddot{x}_1(t) + c_2\ddot{x}_2(t) + \cdots + c_n\ddot{x}_n(t) &= 0 \\ \cdots \quad \cdots \quad \cdots & \\ c_1x_1^{(n-1)}(t) + c_2x_2^{(n-1)}(t) + \cdots + c_nx_n^{(n-1)}(t) &= 0 \end{aligned}$$

Let $t = t_2$ where $t_0 \leqslant t_2 \leqslant t_1$. Then we have the following n simultaneous algebraic equations in $c_1, c_2, \ldots c_n$:

$$\begin{aligned} c_1x_1(t_2) + c_2x_2(t_2) + \cdots + c_nx_n(t_2) &= 0 \\ c_1\dot{x}_1(t_2) + c_2\dot{x}_2(t_2) + \cdots + c_n\dot{x}_n(t_2) &= 0 \\ \cdots \quad \cdots \quad \cdots & \\ c_1x_1^{(n-1)}(t_2) + c_2x_2^{(n-1)}(t_2) + \cdots + c_nx_n^{(n-1)}(t_2) &= 0 \end{aligned}$$

where the $x_i^{(j-1)}(t_2)$ $(i, j = 1, 2, \ldots, n)$ are constants. Since not all of $c_1, c_2, \ldots, c_n$ are zero it follows that

$$\begin{vmatrix} x_1(t_2) & x_2(t_2) & \cdots & x_n(t_2) \\ \dot{x}_1(t_2) & \dot{x}_2(t_2) & \cdots & \dot{x}_n(t_2) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_1^{(n-1)}(t_2) & x_2^{(n-1)}(t_2) & \cdots & x_n^{(n-1)}(t_2) \end{vmatrix} = 0$$

Since t_2 is arbitrary within the time interval $t_0 \leqslant t_2 \leqslant t_1$, this last equation implies that for $t_0 \leqslant t \leqslant t_1$

$$W(x_1, x_2, \ldots, x_n) = \begin{vmatrix} x_1(t) & x_2(t) & \cdots & x_n(t) \\ \dot{x}_1(t) & \dot{x}_2(t) & \cdots & \dot{x}_n(t) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_1^{(n-1)}(t) & x_2^{(n-1)}(t) & \cdots & x_n^{(n-1)}(t) \end{vmatrix} = 0$$

This completes the proof.

Note that the converse of Theorem 6–1 is not necessarily true. For example, consider

$$x_1(t) = t^3, \qquad x_2(t) = |t|^3$$

Then,

$$W(x_1, x_2) = \begin{vmatrix} t^3 & \pm t^3 \\ 3t^2 & \pm 3t^2 \end{vmatrix} = 0$$

Clearly $x_1(t)$ and $x_2(t)$ are linearly independent in the interval $-1 \leqslant t \leqslant 1$.

Theorem 6-2 Suppose that the Wronskian $W(x_1, x_2, \ldots, x_n)$ of n differentiable functions $x_1(t), x_2(t), \ldots, x_n(t)$ is zero for the time interval $t_0 \leqslant t \leqslant t_1$. If Wronskian of $n-1$ functions $x_1(t), x_2(t), \ldots, x_{n-1}(t)$ differs from zero for the entire time interval $t_0 \leqslant t \leqslant t_1$, then the n functions $x_1(t), x_2(t), \ldots, x_n(t)$ are linearly dependent for the time interval $t_0 \leqslant t \leqslant t_1$.

Proof By assumption $W(x_1, x_2, \ldots, x_{n-1}) \neq 0$ for the entire time interval $t_0 \leqslant t \leqslant t_1$. Therefore, for any t between t_0 and t_1 there exists a set of $n-1$ functions of t, that is, $c_1, c_2, \ldots, c_{n-1}$, such that

$$\begin{aligned} c_1x_1(t) + c_2x_2(t) + \cdots + c_{n-1}x_{n-1}(t) &= x_n(t) \\ c_1\dot{x}_1(t) + c_2\dot{x}_2(t) + \cdots + c_{n-1}\dot{x}_{n-1}(t) &= \dot{x}_n(t) \\ \cdots \quad \cdots \quad \cdots & \\ c_1x_1^{(n-2)}(t) + c_2x_2^{(n-2)}(t) + \cdots + c_{n-1}x_{n-1}^{(n-2)}(t) &= x_n^{(n-2)}(t) \end{aligned} \qquad (6\text{–}4)$$

The c_i $(i = 1, 2, \ldots, n-1)$ are differentiable as many times as desired.

Let us put

$$c_1\overset{(n-1)}{x_1}(t) + c_2\overset{(n-1)}{x_2}(t) + \cdots + c_{n-1}\overset{(n-1)}{x_{n-1}}(t) = \overset{(n-1)}{x_n}(t) + c \tag{6-5}$$

By multiplying both sides of the foregoing n equations [Eq. (6–4) and Eq. (6–5)] the cofactors of $x_n, \dot{x}_n, \ldots, \overset{(n-2)}{x_n}, \overset{(n-1)}{x_n}$ in the $n \times n$ determinant $W(x_1, x_2, \ldots, x_n)$, respectively, and summing up the resulting equations we obtain

$$0 = W(x_1, x_2, \ldots, x_n) + cW(x_1, x_2, \ldots, x_{n-1})$$

Since $W(x_1, x_2, \ldots, x_n) = 0$ and $W(x_1, x_2, \ldots, x_{n-1}) \neq 0$, we obtain $c = 0$. Thus Eq. (6–5) becomes

$$c_1\overset{(n-1)}{x_1}(t) + c_2\overset{(n-1)}{x_2}(t) + \cdots + c_{n-1}\overset{(n-1)}{x_{n-1}}(t) = \overset{(n-1)}{x_n}(t) \tag{6-6}$$

To prove that $c_1, c_2, \ldots, c_{n-1}$ are constants, we first differentiate the first of Eq. (6–4) with respect to t and subtract from it the second of Eq. (6–4) and obtain

$$\dot{c}_1x_1(t) + \dot{c}_2x_2(t) + \cdots + \dot{c}_{n-1}x_{n-1}(t) = 0$$

Next, differentiating the second of Eq. (6–4) and subtracting from it the third of Eq. (6–4),

$$\dot{c}_1\dot{x}_1(t) + \dot{c}_2\dot{x}_2(t) + \cdots + \dot{c}_{n-1}\dot{x}_{n-1}(t) = 0$$

Similarly,

$$\dot{c}_1\ddot{x}_1(t) + \dot{c}_2\ddot{x}_2(t) + \cdots + \dot{c}_{n-1}\ddot{x}_{n-1}(t) = 0$$

$$\cdots \qquad \cdots \qquad \cdots$$

$$\dot{c}_1\overset{(n-2)}{x_1}(t) + \dot{c}_2\overset{(n-2)}{x_2}(t) + \cdots + \dot{c}_{n-1}\overset{(n-2)}{x_{n-1}}(t) = 0$$

Hence we obtain

$$\begin{bmatrix} x_1(t) & x_2(t) & \cdots & x_{n-1}(t) \\ \dot{x}_1(t) & \dot{x}_2(t) & \cdots & \dot{x}_{n-1}(t) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \overset{(n-2)}{x_1}(t) & \overset{(n-2)}{x_2}(t) & \cdots & \overset{(n-2)}{x_{n-1}}(t) \end{bmatrix} \begin{bmatrix} \dot{c}_1 \\ \dot{c}_2 \\ \cdot \\ \cdot \\ \cdot \\ \dot{c}_{n-1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

By assumption $W(x_1, x_2, \ldots, x_{n-1}) \neq 0$. Hence

$$\dot{c}_1 = \dot{c}_2 = \cdots = \dot{c}_{n-1} = 0$$

which implies that $c_1, c_2, \ldots, c_{n-1}$ are constants. Therefore, we conclude that $x_1(t), x_2(t), \ldots, x_n(t)$ are linearly dependent for the time interval $t_0 \leqslant t \leqslant t_1$, because, for the time interval $t_0 \leqslant t \leqslant t_1$, we have

$$c_1x_1(t) + c_2x_2(t) + \cdots + c_{n-1}x_{n-1}(t) - x_n(t) = 0$$

which clearly shows that not all of the coefficients of the $x_i(t)$ are zero. [Note that the coefficient of the $x_n(t)$ term is -1.] This completes the proof.

Consider the following scalar differential equation:

$$\overset{(n)}{x} + a_1(t)\overset{(n-1)}{x} + \cdots + a_{n-1}(t)\dot{x} + a_n(t)x = 0$$

By defining

$$x_1 = x$$

$$x_2 = \dot{x}$$

$$\cdots$$

$$x_n = \overset{(n-1)}{x}$$

we obtain

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \cdot \\ \cdot \\ \cdot \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ -a_n(t) & -a_{n-1}(t) & -a_{n-2}(t) & \cdots & -a_1(t) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$

The solutions of this last equation form an n-dimensional vector space. We shall later present a theorem which implies that the necessary and sufficient condition that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ where

$$\mathbf{x}_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \cdot \\ \cdot \\ \cdot \\ x_{ni} \end{bmatrix} = \begin{bmatrix} x_i \\ \dot{x}_i \\ \cdot \\ \cdot \\ \cdot \\ \overset{(n-1)}{x_i} \end{bmatrix}$$

are linearly independent and hence they are a basis of the n-dimensional vector space is that

$$W(t_0) = \begin{vmatrix} x_1(t_0) & x_2(t_0) & \cdots & x_n(t_0) \\ \dot{x}_1(t_0) & \dot{x}_2(t_0) & \cdots & \dot{x}_n(t_0) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \overset{(n-1)}{x_1}(t_0) & \overset{(n-1)}{x_2}(t_0) & \cdots & \overset{(n-1)}{x_n}(t_0) \end{vmatrix} \neq 0$$

(See Theorem 6–4. Since t_0 is arbitrary, if $W(t_0) \neq 0$, then $W(t) \neq 0$ for any t.)

Theorem 6-3 Suppose that $x_1(t), x_2(t), \ldots, x_n(t)$ are n solutions of a linear differential equation

$$L(x) = \overset{(n)}{x} + a_1(t)\overset{(n-1)}{x} + \cdots + a_{n-1}(t)\dot{x} + a_n(t)x = 0$$

where the $a_i(t)$ are absolutely integrable as functions of t in the time interval $t_0 \leqslant t \leqslant t_1$. If the Wronskian $W(x_1, x_2, \ldots, x_n)$ assumes zero at $t = t_2$,

where $t_0 \leqslant t_2 \leqslant t_1$, then $x_1(t), x_2(t), \ldots, x_n(t)$ are linearly dependent in the time interval $t_0 \leqslant t \leqslant t_1$ and $W(x_1, x_2, \ldots, x_n) = 0$ in the same time interval.

Proof By assumption, there exist $c_1, c_2, \ldots, c_n$, not all zero, such that

$$\begin{aligned} c_1x_1(t_2) + c_2x_2(t_2) + \cdots + c_nx_n(t_2) &= 0 \\ c_1\dot{x}_1(t_2) + c_2\dot{x}_2(t_2) + \cdots + c_n\dot{x}_n(t_2) &= 0 \\ \cdots \quad \cdots \quad \cdots \\ \overset{(n-1)}{c_1x_1}(t_2) + \overset{(n-1)}{c_2x_2}(t_2) + \cdots + \overset{(n-1)}{c_nx_n}(t_2) &= 0 \end{aligned}$$

This set of equations implies that the solution

$$y(t) = c_1x_1(t) + c_2x_2(t) + \cdots + c_nx_n(t)$$

of $L(x) = 0$ satisfies

$$y(t_2) = \dot{y}(t_2) = \cdots = \overset{(n-1)}{y}(t_2) = 0$$

On the other hand, the function identically equal to zero is also a solution of $L(x) = 0$, and the coefficients of the derivatives of this function are all zero. By uniqueness of the solution of the linear differential equation, we conclude that $y(t) = 0$. Hence

$$c_1x_1(t) + c_2x_2(t) + \cdots + c_nx_n(t) = 0$$

for the time interval $t_0 \leqslant t \leqslant t_1$. Since not all c's are zero, $x_1(t), x_2(t), \ldots, x_n(t)$ are linearly dependent. By Theorem 6–1, $W(x_1, x_2, \ldots, x_n) = 0$ for $t_0 \leqslant t \leqslant t_1$.

Theorem 6-4 If n solutions $x_1(t), x_2(t), \ldots, x_n(t)$ of a linear differential equation

$$L(x) = \overset{(n)}{x} + a_1(t)\overset{(n-1)}{x} + \cdots + a_{n-1}(t)\dot{x} + a_n(t)x = 0$$

where the $a_i(t)$ are absolutely integrable as functions of t, are linearly independent for the time interval $t_0 \leqslant t \leqslant t_1$, then the Wronskian $W(x_1, x_2, \ldots, x_n)$ is nonzero in the same time interval $t_0 \leqslant t \leqslant t_1$.

Proof The proof follows from Theorems 6–1 and 6–3. Details are left to the reader.

Theorem 6-5 Let n linearly independent solutions of a linear differential equation

$$L(x) = \overset{(n)}{x} + a_1(t)\overset{(n-1)}{x} + \cdots + a_{n-1}(t)\dot{x} + a_n(t)x = 0$$

where the $a_i(t)$ are absolutely integrable as functions of t be $x_1(t), x_2(t), \ldots, x_n(t)$. The general solution of $L(x) = 0$ can be expressed as

$$c_1x_1(t) + c_2x_2(t) + \cdots + c_nx_n(t)$$

where $c_1, c_2, \ldots, c_n$ are constants.

Proof By assumption, $x_1(t), x_2(t), \ldots, x_n(t)$ are linearly independent. By Theorem 6–4, the Wronskian $W(x_1, x_2, \ldots, x_n)$ is nonzero everywhere.

Let an arbitrary solution of $L(x) = 0$ be $z(t)$. Then for an arbitrary time $t = t_1$

$$W(x_1, x_2, \ldots, x_n) \neq 0$$

Therefore, there exist constants $c_1, c_2, \ldots, c_n$, not all zero, such that

$$\begin{aligned} c_1x_1(t_1) + c_2x_2(t_1) + \cdots + c_nx_n(t_1) &= z(t_1) \\ c_1\dot{x}_1(t_1) + c_2\dot{x}_2(t_1) + \cdots + c_n\dot{x}_n(t_1) &= \dot{z}(t_1) \\ \cdots \quad \cdots \quad \cdots \\ \overset{(n-1)}{c_1x_1}(t_1) + \overset{(n-1)}{c_2x_2}(t_1) + \cdots + \overset{(n-1)}{c_nx_n}(t_1) &= \overset{(n-1)}{z}(t_1) \end{aligned}$$

Using these constants $c_1, c_2, \ldots, c_n$ put

$$y(t) = c_1x_1(t) + c_2x_2(t) + \cdots + c_nx_n(t)$$

Then

$$L(y) = 0$$

and

$$y(t_1) = z(t_1), \qquad \dot{y}(t_1) = \dot{z}(t_1), \quad \ldots, \quad \overset{(n-1)}{y}(t_1) = \overset{(n-1)}{z}(t_1)$$

Thus $y(t)$ and $z(t)$ are solutions of the given differential equation satisfying the same initial conditions. By uniqueness of the solution of the linear differential equation, $y(t) \equiv z(t)$. Hence we have proved that an arbitrary solution $z(t)$ can be expressed as

$$z(t) = c_1x_1(t) + c_2x_2(t) + \cdots + c_nx_n(t).$$

Since $z(t)$ involves n arbitrary constants, it is a general solution. This completes the proof.

Fundamental matrix Let us form an $n \times n$ matrix $\mathbf{X}$ such that its n columns consist of n linearly independent solutions, or n basis vectors of Eq. (6–2). Such a matrix $\mathbf{X}$ is called a *fundamental matrix*. A fundamental matrix is nonsingular. Any fundamental matrices satisfy the matrix differential equation $\dot{\mathbf{X}} = \mathbf{A}(t)\mathbf{X}$. Notice that fundamental matrices differ from each other by a multiplicative constant matrix. For a given initial condition, the fundamental matrix is uniquely determined.

Let an $n \times n$ matrix $\Phi(t, t_0)$ be the solution of the following matrix differential equation:

$$\frac{d}{dt}\Phi(t, t_0) = \mathbf{A}(t)\Phi(t, t_0), \qquad \Phi(t_0, t_0) = \mathbf{I} \tag{6-7}$$

where $\mathbf{A}(t)$ is the same $n \times n$ matrix defined in Eq. (6-2). Any fundamental matrix $\mathbf{X}(t)$ of Eq. (6-2) can be written as

$$\mathbf{X}(t) = \Phi(t, t_0)\mathbf{C}$$

where $\mathbf{C}$ is a nonsingular matrix. Let us denote n linearly independent column vectors of $\Phi(t, t_0)$ as $\Phi_1(t, t_0)$, $\Phi_2(t, t_0)$, . . . , $\Phi_n(t, t_0)$. Then the $\Phi_i(t, t_0)$ define a set of basis vectors.

Notice that

$$\Phi_i(t, t_0) = \begin{bmatrix} \phi_{1i}(t, t_0) \\ \phi_{2i}(t, t_0) \\ \cdot \\ \cdot \\ \cdot \\ \phi_{ni}(t, t_0) \end{bmatrix} \qquad (i = 1, 2, \ldots, n)$$

represents the response of the system with the initial condition given by

$$\Phi_i(t_0, t_0) = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \Big\} i - 1$$

The basis vectors $\Phi_i(t, t_0)$ do not involve delta functions and are infinitely differentiable for $t > t_0$.

Note that if the $\Phi_i(t, t_0)$ $(i = 1, 2, \ldots, n)$ constitute a set of basis vectors of a given system and if the $\hat{\Phi}_i(t, t_0)$ $(i = 1, 2, \ldots, n)$ constitute a set of basis vectors of the same system, then

$$\Phi(t, t_0) = \hat{\Phi}(t, t_0)\mathbf{C}$$

where $\mathbf{C}$ is a nonsingular matrix. Note also that

$$\Phi(t, t_0) = \mathbf{X}(t)\mathbf{X}^{-1}(t_0)$$

Solution of homogeneous vector matrix differential equation

Consider next the homogeneous linear vector matrix differential equation

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t), \qquad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{6-8}$$

where $\mathbf{x}(t)$ is an n vector and $\mathbf{A}(t)$ is the same $n \times n$ matrix defined in Eq. (6-2). Any solution of Eq. (6-8) is given by

$$\mathbf{x}(t) = \Phi(t, t_0)\mathbf{x}(t_0) \tag{6-9}$$

where $\Phi(t, t_0)$ is the unique solution of Eq. (6-7). We can easily verify this. Since

$$\mathbf{x}(t_0) = \Phi(t_0, t_0)\,\mathbf{x}(t_0) = \mathbf{I}\mathbf{x}_0 = \mathbf{x}_0$$

and

$$\dot{\mathbf{x}}(t) = \frac{d}{dt}[\Phi(t, t_0)\mathbf{x}(t_0)] = \frac{d}{dt}[\Phi(t, t_0)]\mathbf{x}(t_0)$$
$$= \mathbf{A}(t)\Phi(t, t_0)\mathbf{x}(t_0) = \mathbf{A}(t)\mathbf{x}(t)$$

we have verified that Eq. (6-9) is the solution of Eq. (6-8).

We see that Eq. (6-9), the homogeneous solution of Eq. (6-8), is simply a transformation of the initial condition. Hence the unique fundamental matrix $\Phi(t, t_0)$ is also called the *state transition matrix*. By Eq. (6-9) it maps the state $\mathbf{x}(t_0)$ (or $\mathbf{x}_0$) at time t_0 into the state $\mathbf{x}(t)$ at time t. The state transition matrix contains all informations about the property of free motions of the system defined by Eq. (6-8).

The unique solution $\mathbf{x}(t)$ observed at time t after starting at time t_0 in state $\mathbf{x}_0$ is sometimes written as $\phi(t; \mathbf{x}_0, t_0)$. As time elapses $\phi(t; \mathbf{x}_0, t_0)$ represents a trajectory in the n dimensional vector space. Using this notation, if $\mathbf{x}(t_0) = \mathbf{x}_0$, $\phi(t_2; \mathbf{x}_0, t_0) = \mathbf{x}_2$, and $\phi(t_3; \mathbf{x}_0, t_0) = \mathbf{x}_3$, then we have $\phi(t_3; \mathbf{x}_0, t_0) = \phi(t_3; \mathbf{x}_2, t_2) = \mathbf{x}_3$. Hence

$$\mathbf{x}_3 = \phi(t_3; \mathbf{x}_2, t_2) = \Phi(t_3, t_2)\mathbf{x}_2 = \Phi(t_3, t_2)\,\Phi(t_2, t_0)\mathbf{x}_0$$

Also

$$\mathbf{x}_3 = \phi(t_3; \mathbf{x}_2, t_2) = \phi(t_3; \mathbf{x}_0, t_0) = \Phi(t_3, t_0)\mathbf{x}_0$$

Hence

$$\Phi(t_3, t_2)\,\Phi(t_2, t_0) = \Phi(t_3, t_0)$$

Since t_3 may be any time within the interval $t_0 \leqslant t_3 \leqslant t_1$, upon substituting $t_3 = t_0$ in this last equation we obtain

$$\Phi(t_0, t_2)\,\Phi(t_2, t_0) = \mathbf{I}$$

Since t_2 can be arbitrary within the interval $t_0 \leqslant t_2 \leqslant t_1$, dropping the subscript 2 we obtain

$$\Phi(t_0, t)\,\Phi(t, t_0) = \mathbf{I} \qquad \text{or} \qquad \Phi^{-1}(t, t_0) = \Phi(t_0, t)$$

Example 6-1 Consider the following differential equation:

$$\ddot{x} + 2\dot{x} = 0$$

This equation can be written as

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = -2x_2 \tag{6-10}$$

where $x_1 = x(t)$ and $x_2 = \dot{x}(t)$. Rewriting in the form of the vector matrix differential equation, we have

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) \tag{6-11}$$

where

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & -2 \end{bmatrix}$$

We shall assume that $t_0 = 0$. The solution of Eq. (6–10) can be obtained by use of an elementary technique as follows:

$$\begin{aligned} x(t) &= x_1(t) = c_0 + c_1 e^{-2t} \\ \dot{x}(t) &= x_2(t) = -2c_1 e^{-2t} \end{aligned} \tag{6-12}$$

where c_0 and c_1 are arbitrary constants. From Eq. (6–12) it is possible to write

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = c_0 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + c_1 \begin{bmatrix} e^{-2t} \\ -2e^{-2t} \end{bmatrix}$$

or

$$\mathbf{x}(t) = c_0\mathbf{x}_1(t) + c_1\mathbf{x}_2(t)$$

where

$$\mathbf{x}_1(t) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \mathbf{x}_2(t) = \begin{bmatrix} e^{-2t} \\ -2e^{-2t} \end{bmatrix}$$

In the following, we shall see that $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ are solutions of Eq. (6–11) and are linearly independent of each other. Since

$$\dot{\mathbf{x}}_1(t) = \frac{d}{dt}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -2 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \mathbf{A}\mathbf{x}_1(t)$$

and

$$\dot{\mathbf{x}}_2(t) = \frac{d}{dt}\begin{bmatrix} e^{-2t} \\ -2e^{-2t} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -2 \end{bmatrix}\begin{bmatrix} e^{-2t} \\ -2e^{-2t} \end{bmatrix} = \mathbf{A}\mathbf{x}_2(t)$$

$\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ are solutions of Eq. (6–11). Linear independence of $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ may be seen from the following: if

$$c_0\mathbf{x}_1(t) + c_1\mathbf{x}_2(t) = \mathbf{0}$$

which means

$$\begin{aligned} c_0 + c_1 e^{-2t} &= 0 \\ -2c_1 e^{-2t} &= 0 \end{aligned}$$

then it is obvious that

$$c_0 = c_1 = 0$$

Thus, $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ are linearly independent. Since $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ are solutions of Eq. (6–11) and are linearly independent of each other, $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ form a fundamental matrix of Eq. (6–11). A few examples of fundamental matrices of Eq. (6–11) are

$$\begin{bmatrix} 1 & e^{-2t} \\ 0 & -2e^{-2t} \end{bmatrix}, \quad \begin{bmatrix} 1 & 1-e^{-2t} \\ 0 & 2e^{-2t} \end{bmatrix}, \quad \begin{bmatrix} 1+e^{-2t} & 1-e^{-2t} \\ -2e^{-2t} & 2e^{-2t} \end{bmatrix}, \quad \text{etc.}$$

For a given initial condition, the fundamental matrix is uniquely determined. Noting that in the present analysis $t_0 = 0$, we obtain the fundamental matrix $\Phi(t, t_0) \equiv \Phi(t)$ which satisfies $\Phi(0) = \mathbf{I}$ as

$$\begin{bmatrix} 1 & \frac{1}{2}(1-e^{-2t}) \\ 0 & e^{-2t} \end{bmatrix}$$

Hence

$$\Phi_1(t) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \Phi_2(t) = \begin{bmatrix} \frac{1}{2}(1-e^{-2t}) \\ e^{-2t} \end{bmatrix}$$

It can be verified that fundamental matrices differ from each other by a multiplicative constant matrix as follows:

$$\begin{bmatrix} 1 & 1-e^{-2t} \\ 0 & 2e^{-2t} \end{bmatrix} = \begin{bmatrix} 1 & e^{-2t} \\ 0 & -2e^{-2t} \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix}$$

$$\begin{bmatrix} 1+e^{-2t} & 1-e^{-2t} \\ -2e^{-2t} & 2e^{-2t} \end{bmatrix} = \begin{bmatrix} 1 & e^{-2t} \\ 0 & -2e^{-2t} \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & e^{-2t} \\ 0 & -2e^{-2t} \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{2}(1-e^{-2t}) \\ 0 & e^{-2t} \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & -2 \end{bmatrix}$$

We point out an important property of $\Phi(t)$. For the constant matrix $\mathbf{A}$ we have

$$\Phi(t+\tau) = \Phi(t)\Phi(\tau) = \Phi(\tau)\Phi(t)$$

from which

$$\Phi(t) = \Phi(\tau)\Phi(t-\tau)$$

or

$$\Phi_i(t) = \Phi(\tau)\Phi_i(t-\tau) \qquad (i = 1, 2, \ldots, n)$$

where the Φ_i are the n columns of Φ. Thus $\Phi_i(t)$ can be expressed as a transformation of $\Phi_i(t-\tau)$.

In the system considered in Example 6–1, $\Phi(t)$ is given by

$$\Phi(t) = \begin{bmatrix} 1 & \frac{1}{2}(1-e^{-2t}) \\ 0 & e^{-2t} \end{bmatrix}$$

Hence

$$\Phi(t-\tau) = \begin{bmatrix} 1 & \frac{1}{2}(1-e^{-2(t-\tau)}) \\ 0 & e^{-2(t-\tau)} \end{bmatrix}$$

The foregoing property can be verified thus:

$$\begin{aligned} \Phi_1(t) &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{2}(1-e^{-2\tau}) \\ 0 & e^{-2\tau} \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &= \Phi(\tau)\Phi_1(t-\tau) \end{aligned}$$

also

$$\begin{aligned} \Phi_2(t) &= \begin{bmatrix} \frac{1}{2}(1-e^{-2t}) \\ e^{-2t} \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{2}(1-e^{-2\tau}) \\ 0 & e^{-2\tau} \end{bmatrix}\begin{bmatrix} \frac{1}{2}(1-e^{-2(t-\tau)}) \\ e^{-2(t-\tau)} \end{bmatrix} \\ &= \Phi(\tau)\Phi_2(t-\tau) \end{aligned}$$

If $\mathbf{A}(t)$ is a constant matrix, then $\Phi(t, t_0)$ can be obtained as an exponential function of $\mathbf{A}t$. (Refer to Section 6–3.) This is not, however, the case for the time-varying matrix $\mathbf{A}(t)$, since $\exp \int_{t_0}^{t} \mathbf{A}(\tau)\,d\tau$ is not necessarily the fundamental matrix $\Phi(t, t_0)$ unless $\mathbf{A}(t)$ and $\int_{t_0}^{t} \mathbf{A}(\tau)d\tau$ commute. Namely,

$$\Phi(t, t_0) \neq \exp \int_{t_0}^{t} \mathbf{A}(\tau)d\tau \qquad \begin{pmatrix}\text{if } \mathbf{A}(t) \text{ and } \int_{t_0}^{t} \mathbf{A}(\tau)d\tau \\ \text{do not commute for all } t\end{pmatrix}$$

and

$$\Phi(t, t_0) = \exp \int_{t_0}^{t} \mathbf{A}(\tau)d\tau \qquad \begin{pmatrix}\text{if } \mathbf{A}(t) \text{ and } \int_{t_0}^{t} \mathbf{A}(\tau)d\tau \\ \text{commute for all } t\end{pmatrix}$$

Notice that if $\mathbf{A}$ is a constant matrix or if $\mathbf{A}(t)$ is a diagonal matrix, then clearly $\mathbf{A}(t)$ and $\int_{t_0}^{t} \mathbf{A}(\tau)\,d\tau$ commute. For a time-varying matrix $\mathbf{A}(t)$ there is no convenient formula for computing $\Phi(t, t_0)$ and, thererfore, except for simple cases we frequently have to resort to numerical integrations for computing $\Phi(t, t_0)$ in time-varying systems.

Solution of nonhomogeneous vector matrix differential equation Consider the following vector matrix differential equation:

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t), \qquad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{6–13}$$

where $\mathbf{x}(t)$ is an n vector; $\mathbf{u}(t)$, an r vector; $\mathbf{A}(t)$, an $n \times n$ matrix; and $\mathbf{B}(t)$, an $n \times r$ matrix. Here we assume that the elements of $\mathbf{A}(t)$ and those of $\mathbf{B}(t)$ are absolutely integrable as functions of t in the interval $t_0 \leqslant t \leqslant t_1$. Let $\Phi(t, t_0)$ be the unique matrix which satisfies Eq. (6–7), rewritten as follows:

$$\frac{d}{dt}\Phi(t, t_0) = \mathbf{A}(t)\Phi(t, t_0), \qquad \Phi(t_0, t_0) = \mathbf{I}$$

Assume that

$$\mathbf{x}(t) = \Phi(t, t_0)\mathbf{y}(t)$$

then

$$\dot{\mathbf{x}} = \frac{d}{dt}(\Phi\mathbf{y}) = \Phi\dot{\mathbf{y}} + \dot{\Phi}\mathbf{y} = \Phi\dot{\mathbf{y}} + \mathbf{A}\Phi\mathbf{y} = \mathbf{A}\Phi\mathbf{y} + \mathbf{B}\mathbf{u}$$

Therefore,

$$\Phi\dot{\mathbf{y}} = \mathbf{B}\mathbf{u}$$

or

$$\mathbf{y}(t) = \int_{t_0}^{t} \Phi^{-1}(\tau, t_0)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau + \mathbf{y}(t_0)$$

Since

$$\mathbf{y}(t_0) = \Phi^{-1}(t_0, t_0)\mathbf{x}(t_0) = \mathbf{x}(t_0) = \mathbf{x}_0$$

the solution of Eq. (6–13) is given by

$$\begin{aligned}\mathbf{x}(t) &= \Phi(t, t_0)\mathbf{y}(t) \\ &= \Phi(t, t_0)\mathbf{x}_0 + \Phi(t, t_0)\int_{t_0}^{t} \Phi^{-1}(\tau, t_0)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau\end{aligned} \tag{6–14}$$

This solution can also be written as

$$\mathbf{x}(t) = \Phi(t, t_0)\mathbf{x}_0 + \int_{t_0}^{t} \Phi(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau$$

Notice that the solution of the vector matrix differential equation (6–13) is made up of the homogeneous solution $\Phi(t, t_0)\mathbf{x}_0$, and a particular solution $\mathbf{x}_p(t)$ given by

$$\begin{aligned}\mathbf{x}_p(t) &= \Phi(t, t_0)\int_{t_0}^{t} \Phi^{-1}(\tau, t_0)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau \\ &= \int_{t_0}^{t} \Phi(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau\end{aligned}$$

6-3 SOLUTIONS OF LINEAR TIME-INVARIANT VECTOR MATRIX DIFFERENTIAL EQUATIONS

Although most physical systems are more or less time-varying, many of them may be represented approximately by linear time-invariant equations. This approximation is good if characteristics of the system change very slowly compared with the input variations. In this section we shall give a particular attention to linear time-invariant differential equation systems. The general method for finding solutions of linear vector matrix differential equations presented in Section 6–2 can be simplified for time-invariant systems.

Notice that in time-invariant systems the response does not depend on t_0, the time the input is applied. Hence we can conveniently choose t_0 as the origin of the time axis and consider that t_0 equals 0 in order to simplify presentation of the analysis of linear time-invariant systems. We shall therefore put $t_0 = 0$ in the following analyses, unless otherwise stated.

Homogeneous equation (free dynamic system) Consider the free motion of the dynamic system described by

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t), \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{6–15}$$

where $\mathbf{x}(t)$ is an n vector which defines the state of the system at time t and $\mathbf{A}$ is an $n \times n$ constant matrix. The solution of this vector matrix differential equation can be obtained by various methods. One method was presented in Section 3–8. We shall first review this method (which we shall call Method

1) and shall present two other methods (Method 2 and Method 3) in what follows:

Method 1 In Section 3–8, the system equation, Eq. (6–15), was transformed by the transformation $\mathbf{x} = \mathbf{P}\mathbf{y}$ into

$$\dot{\mathbf{y}} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}\mathbf{y}$$

If A has distinct eigenvectors, then $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ can be made equal to $\mathbf{D} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and the solution is found as

$$\mathbf{y}(t) = \begin{bmatrix} e^{\lambda_1 t} & & & 0 \\ & e^{\lambda_2 t} & & \\ & & \ddots & \\ 0 & & & e^{\lambda_n t} \end{bmatrix} \mathbf{y}(0) = \mathbf{Q}(t)\mathbf{y}(0)$$

Since

$$\mathbf{Q}(t) = \begin{bmatrix} e^{\lambda_1 t} & & & 0 \\ & e^{\lambda_2 t} & & \\ & & \ddots & \\ 0 & & & e^{\lambda_n t} \end{bmatrix} = e^{\mathbf{D}t}$$

we obtain

$$\mathbf{y}(t) = e^{\mathbf{D}t}\mathbf{y}(0)$$

or

$$\begin{aligned} \mathbf{x}(t) &= \mathbf{P}\mathbf{Q}(t)\mathbf{P}^{-1}\mathbf{x}(0) \\ &= \mathbf{P}e^{\mathbf{D}t}\mathbf{P}^{-1}\mathbf{x}(0) \\ &= \mathbf{P}e^{(\mathbf{P}^{-1}\mathbf{A}\mathbf{P})t}\mathbf{P}^{-1}\mathbf{x}(0) \end{aligned} \tag{6–16}$$

In the case where A involves multiple eigenvalues λ_i of multiplicity m_i, $(i = 1, 2, \ldots, k;\ m_1 + m_2 + \ldots + m_k = n)$ and the eigenvectors corresponding to λ_i are also multiple with the corresponding multiplicity m_i, then $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ becomes the Jordan canonical form, which we shall denote by J. The solution $\mathbf{x}(t)$ in this case is given by

$$\begin{aligned} \mathbf{x}(t) &= \mathbf{P}\mathbf{S}(t)\mathbf{Q}(t)\mathbf{P}^{-1}\mathbf{x}(0) \\ &= \mathbf{P}e^{\mathbf{J}t}\mathbf{P}^{-1}\mathbf{x}(0) \\ &= \mathbf{P}e^{(\mathbf{P}^{-1}\mathbf{A}\mathbf{P})t}\mathbf{P}^{-1}\mathbf{x}(0) \end{aligned} \tag{6–17}$$

where $\mathbf{S}(t)$ is given by

$$\mathbf{S}(t) = \begin{bmatrix} \mathbf{S}_1(t) & & & 0 \\ & \mathbf{S}_2(t) & & \\ & & \ddots & \\ 0 & & & \mathbf{S}_k(t) \end{bmatrix}$$

where

$$\mathbf{S}_i(t) = \begin{bmatrix} 1 & t & \frac{t^2}{2!} & \cdots & \frac{t^{m_i-1}}{(m_i-1)!} \\ 0 & 1 & t & \cdots & \frac{t^{m_i-2}}{(m_i-2)!} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & t \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} = m_i \times m_i \text{ matrix}$$

[Refer to Section 3–8 for derivation of $\mathbf{S}_i(t)$.]

Method 2 Another way to find the solution of Eq. (6–15) is to proceed as follows: By analogy with the scalar case

$$\dot{\xi}(t) = a\xi(t), \qquad \xi(0) = \xi_0$$

whose solution is $\xi(t) = e^{at}\xi_0$, we use the matrix exponential

$$e^{\mathbf{A}t} = \mathbf{I} + \sum_{k=1}^{\infty} \frac{\mathbf{A}^k t^k}{k!}$$

for finding the solution of Eq. (6–15). The series $e^{\mathbf{A}t}$ converges absolutely and uniformly in any finite interval of the time axis. Since

$$\frac{d}{dt}(e^{\mathbf{A}t}) = \mathbf{A}e^{\mathbf{A}t}$$

the solution of Eq. (6–15) satisfying the initial condition $\mathbf{x}(t_0) = \mathbf{x}_0$ is given by

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}_0 \tag{6–18}$$

For $t = 0$, $e^{\mathbf{A}t}$ reduces to the identity matrix. Therefore, Eq. (6–18) clearly satisfies the initial condition.

In this approach we must compute $e^{\mathbf{A}t}$ for a given A. Computation of $e^{\mathbf{A}t}$ can be accomplished by a few different ways. (Four such methods are presented in Section 6–4.)

Method 3 The third method is based on the analog computer simulation for solving the vector matrix differential equation. Essentially the simulation consists of n integrators. As an example, consider the following system:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Figure 6–1 shows an analog computer diagram for solving this third-order system. In order to simplify the essential part of the diagram, we made the convenient assumption that the a_{ij} are negative and $|a_{ij}| \leqslant 1$. This assumption is, of course, inessential and the diagram can be modified easily to the general case where the a_{ij} assume any real values.

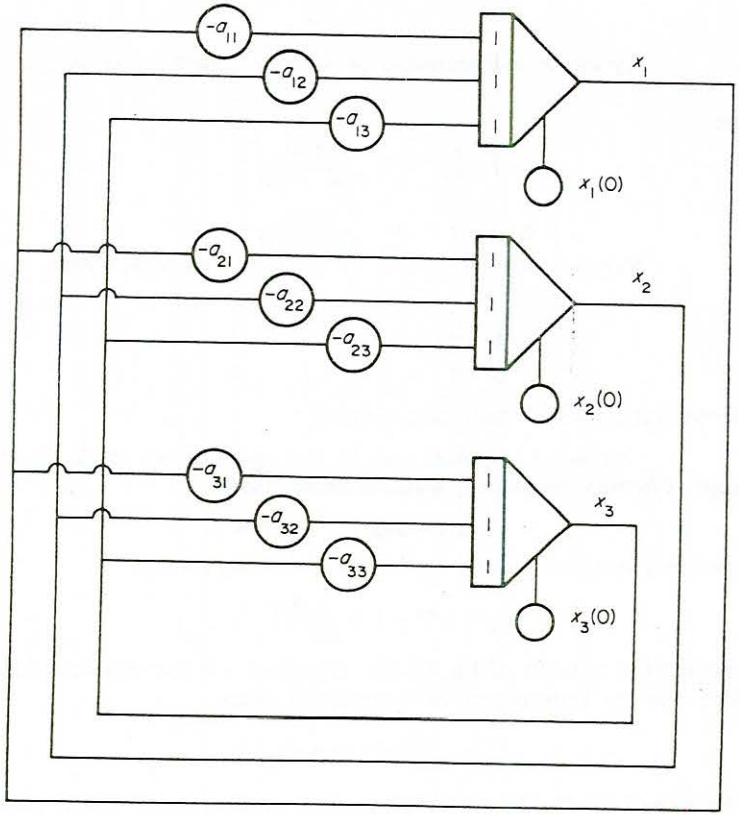


Figure 6-1 Analog computer diagram for solving $\dot{\mathbf{x}} = \mathbf{Ax}$.

A few comments on $\Phi(t)$[1] From Eq. (6–16), Eq. (6–17), and Eq. (6–18) we note that the unique fundamental matrix $\Phi(t)$ which satisfies

$$\dot{\Phi}(t) = \mathbf{A}\Phi(t), \qquad \Phi(0) = \mathbf{I}$$

is given by

$$\begin{aligned}\Phi(t) &= e^{\mathbf{A}t} \\ &= \mathbf{P}e^{\mathbf{D}t}\mathbf{P}^{-1} \qquad (\text{or } \mathbf{P}e^{\mathbf{J}t}\mathbf{P}^{-1}) \\ &= \mathbf{P}e^{(\mathbf{P}^{-1}\mathbf{AP})t}\mathbf{P}^{-1}\end{aligned} \tag{6–19}$$

Notice that since

$$e^{\mathbf{A}t} = \mathbf{P}e^{(\mathbf{P}^{-1}\mathbf{AP})t}\mathbf{P}^{-1}$$

we have

$$\mathbf{P}^{-1}e^{\mathbf{A}t}\mathbf{P} = e^{(\mathbf{P}^{-1}\mathbf{AP})t}$$

Notice also that, for an $n \times n$ constant matrix $\mathbf{A}$,

$$\Phi(t+\tau) = e^{\mathbf{A}(t+\tau)} = \Phi(t)\Phi(\tau) = \Phi(\tau)\Phi(t) \qquad t, \tau \geqslant 0$$

and

$$\Phi^{-1}(t) = e^{-\mathbf{A}t} = \Phi(-t)$$

Hence the solution of

$$\dot{\mathbf{x}} = \mathbf{Ax}, \qquad \mathbf{x}(t_0) = \mathbf{x}_0$$

can be written as

$$\mathbf{x}(t) = e^{\mathbf{A}(t-t_0)}\mathbf{x}(t_0) = \Phi(t-t_0)\mathbf{x}(t_0)$$

From Eq. (6–19) we see that the elements of $\Phi(t)$ consist of terms like $t^h e^{\lambda_i t}$ ($h = 0, 1, 2, \ldots, m_i - 1$, where m_i is the multiplicity of the eigenvector corresponding to the eigenvalue λ_i). Specifically, if the eigenvectors of $\mathbf{A}$ are distinct, elements of $\Phi(t)$ consist of $e^{\lambda_i t}$ ($i = 1, 2, \ldots, n$). If the eigenvectors of $\mathbf{A}$ involve, for example, one multiple eigenvector corresponding to the eigenvalue λ_1, and $n - m$ distinct eigenvectors corresponding to $\lambda_{m+1}, \ldots, \lambda_n$, then the elements of $\Phi(t)$ will involve the terms $t^h e^{\lambda_i t}$ ($h = 0, 1, 2, \ldots, m-1$) and $e^{\lambda_i t}$ ($i = m+1, \ldots, n$). Hence the elements of $\Phi(t)$ can be differentiated as many times as we wish for $t > 0$.

Nonhomogeneous equation (forced dynamic system) Consider the following vector matrix differential equation:

$$\dot{\mathbf{x}}(t) = \mathbf{Ax}(t) + \mathbf{Bu}(t), \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{6–20}$$

where $\mathbf{x}$ is an n vector, $\mathbf{u}$ is an r vector, $\mathbf{A}$ is an $n \times n$ constant matrix, and $\mathbf{B}$ is an $n \times r$ constant matrix. By writing Eq. (6–20) as

$$\dot{\mathbf{x}}(t) - \mathbf{Ax}(t) = \mathbf{Bu}(t)$$

and multiplying both sides of this equation by $e^{-\mathbf{A}t}$, we obtain

$$e^{-\mathbf{A}t}[\dot{\mathbf{x}}(t) - \mathbf{Ax}(t)] = \frac{d}{dt}[e^{-\mathbf{A}t}\mathbf{x}(t)] = e^{-\mathbf{A}t}\mathbf{Bu}(t)$$

Integrating the preceding equation between 0 and t we get

$$e^{-\mathbf{A}t}\mathbf{x}(t) = \mathbf{x}_0 + \int_0^t e^{-\mathbf{A}\tau}\mathbf{Bu}(\tau)\,d\tau$$

or

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}_0 + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{Bu}(\tau)\,d\tau \tag{6–21}$$

This is the solution to Eq. (6–20). Of course, the same solution can be obtained from Eq. (6–14). To demonstrate this, rewrite Eq. (6–14) as follows:

$$\mathbf{x}(t) = \Phi(t)\mathbf{x}_0 + \Phi(t)\int_0^t \Phi^{-1}(\tau)\mathbf{Bu}(\tau)\,d\tau \tag{6–22}$$

[1]For time-invariant (or constant coefficient) systems $\Phi(t, t_0) = \Phi(t - t_0, 0)$. Hence for such systems we usually use the simplified notation $\Phi(t - t_0)$ or $\Phi(t)$ if $t_0 = 0$, rather than $\Phi(t, t_0)$ or $\Phi(t - t_0, 0)$.

where $\Phi(t)$ is given as

$$\Phi(t) = e^{\mathbf{A}t}$$

Since

$$\Phi(t)\Phi^{-1}(\tau) = \Phi(t-\tau) = e^{\mathbf{A}(t-\tau)}$$

Eq. (6–22) can be simplified into

$$\mathbf{x}(t) = \Phi(t)\mathbf{x}_0 + \int_0^t \Phi(t-\tau)\mathbf{B}\mathbf{u}(\tau)\,d\tau$$

$$= e^{\mathbf{A}t}\mathbf{x}_0 + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)\,d\tau$$

which is the same as Eq. (6–21).

Notice that if the initial time is given as t_0 instead of 0, then we have

$$\mathbf{x}(t) = e^{\mathbf{A}(t-t_0)}\mathbf{x}_0 + \int_{t_0}^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)\,d\tau$$

where

$$\mathbf{x}_0 = \mathbf{x}(t_0)$$

The response $\mathbf{x}(t)$ is a sum of the motions due to the initial condition and those due to the forcing function. The motions due to the forcing function

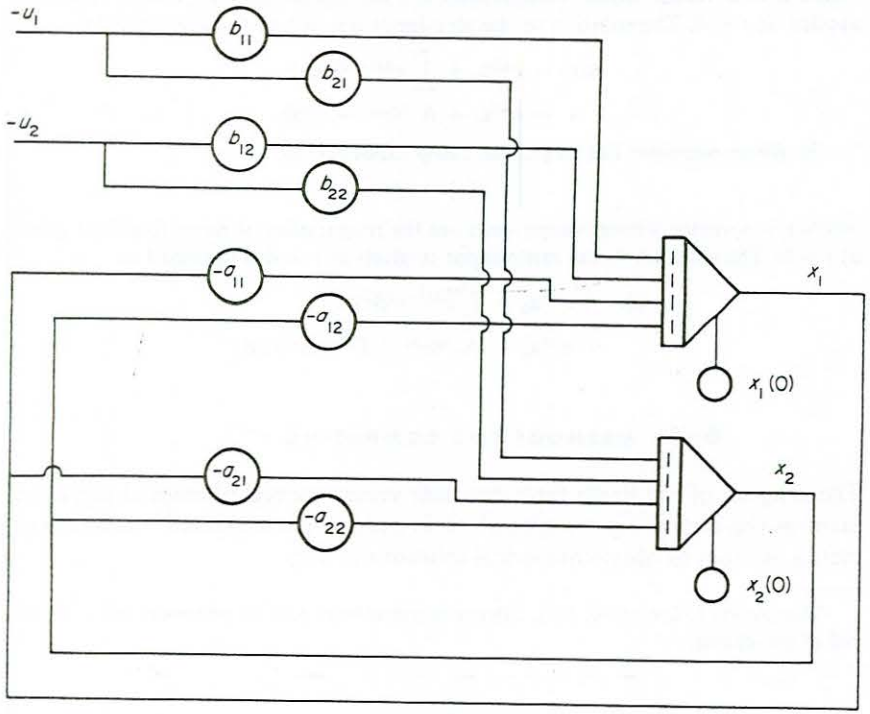


Figure 6-2 Analog computer diagram for solving $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$.

depend on **B**. It is possible to choose **B** such that any particular motion $e^{\lambda_i t}$ cannot be excited for any input. This particular motion can be excited only by the initial condition. (Refer to Example 6–3.)

Next we consider an analog computer simulation of the nonhomogeneous vector matrix differential equation. Consider the following system:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

The solution of this differential equation can be obtained by the analog computer simulation shown in Fig. 6–2. In the diagram we assumed that the values of the a_{ij} are negative and $|a_{ij}| \leqslant 1$; the values of the b_{ij} are positive and $|b_{ij}| \leqslant 1$. Such assumptions, however, are inessential and the diagram can be easily modified to handle the general case where the a_{ij} and the b_{ij} assume any real values.

Example 6-2 Obtain the solution of the following vector matrix differential equation where u is a unit step function occurring at $t = 0$:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} -6 & 1 & 0 \\ -11 & 0 & 1 \\ -6 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 2 \\ 6 \\ 2 \end{bmatrix} [u] \tag{6–23}$$

or

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u$$

Referring to Example 4–2, we find that by the following transformation

$$\mathbf{x} = \mathbf{P}\mathbf{y}$$

or

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -1 & 2 & 1 \\ -5 & 8 & 3 \\ -6 & 6 & 2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

Eq. (6–23) can be transformed into

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} [u]$$

or

$$\dot{\mathbf{y}} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}\mathbf{y} + \mathbf{P}^{-1}\mathbf{B}u \tag{6–24}$$

The solution of Eq. (6–24) is obtained as follows:

$$\mathbf{y}(t) = e^{\mathbf{P}^{-1}\mathbf{A}\mathbf{P}t}\mathbf{y}_0 + \int_0^t e^{\mathbf{P}^{-1}\mathbf{A}\mathbf{P}(t-\tau)}\mathbf{P}^{-1}\mathbf{B}u\,d\tau$$

or

$$\begin{bmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \end{bmatrix} = \begin{bmatrix} e^{-t} & 0 & 0 \\ 0 & e^{-2t} & 0 \\ 0 & 0 & e^{-3t} \end{bmatrix} \begin{bmatrix} y_1(0) \\ y_2(0) \\ y_3(0) \end{bmatrix} + \begin{bmatrix} 1 - e^{-t} \\ \frac{1}{2}(1 - e^{-2t}) \\ \frac{1}{3}(1 - e^{-3t}) \end{bmatrix}$$

In terms of the original state vector $\mathbf{x}(t)$, the solution is given by

$$\mathbf{x}(t) = \mathbf{P}e^{\mathbf{P}^{-1}\mathbf{A}\mathbf{P}t}\mathbf{P}^{-1}\mathbf{x}_0 + \mathbf{P}\int_0^t e^{\mathbf{P}^{-1}\mathbf{A}\mathbf{P}(t-\tau)}\mathbf{P}^{-1}\mathbf{B}u\,d\tau$$

or

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} = \begin{bmatrix} -1 & 2 & 1 \\ -5 & 8 & 3 \\ -6 & 6 & 2 \end{bmatrix} \begin{bmatrix} e^{-t} & 0 & 0 \\ 0 & e^{-2t} & 0 \\ 0 & 0 & e^{-3t} \end{bmatrix} \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ -2 & 1 & -\frac{1}{2} \\ \frac{9}{2} & -\frac{5}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} x_1(0) \\ x_2(0) \\ x_3(0) \end{bmatrix}$$
$$+ \begin{bmatrix} -1 & 2 & 1 \\ -5 & 8 & 3 \\ -6 & 6 & 2 \end{bmatrix} \begin{bmatrix} 1 - e^{-t} \\ \frac{1}{2}(1 - e^{-2t}) \\ \frac{1}{3}(1 - e^{-3t}) \end{bmatrix}$$

Example 6-3 In this example we demonstrate that in the system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u, \qquad \mathbf{x}(0) = \mathbf{x}_0$$

if $\mathbf{B}$ assumes a certain form, then a particular motion $e^{\lambda_i t}$ cannot be excited for any input. Such a particular motion can be excited only by the initial condition.

Consider the system described by

$$(p + 1)(p + 2)(p + 3)x = (p + 1)u, \qquad p = \frac{d}{pt}$$

Referring to Example 4–4, the simplified state space equation of this system is

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} [u]$$

The solution of this last equation is given by

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} e^{-t} & 0 & 0 \\ 0 & e^{-2t} & 0 \\ 0 & 0 & e^{-3t} \end{bmatrix} \begin{bmatrix} y_1(0) \\ y_2(0) \\ y_3(0) \end{bmatrix} + \int_0^t \begin{bmatrix} 0 \\ e^{-2(t-\tau)} \\ e^{-3(t-\tau)} \end{bmatrix} u(\tau)\,d\tau$$

Clearly, the forcing function u is uncoupled to the motion e^{-t}. (The first component of the control matrix in the simplified state space representation is zero.) Therefore, this motion can be excited only by the nonzero initial condition $y_1(0)$. This situation occurs whenever factors in the numerator and denominator of the transfer function cancel each other. (In the present system the terms $(p + 1)$ in both sides of the equation cancel each other.) We shall discuss such cases again in Section 7–6.

Example 6-4 Consider the system described by

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \qquad \mathbf{x}(0-) = \mathbf{x}_0$$

where $\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector or input vector)
$\mathbf{A} = n \times n$ matrix
$\mathbf{B} = n \times r$ matrix

Obtain the response to each of the following three inputs:

1. The r components of $\mathbf{u}$ are impulse functions of various magnitudes.
2. The r components of $\mathbf{u}$ are step functions of various magnitudes.
3. The r components of $\mathbf{u}$ are ramp functions of various magnitudes.

Assume that each input is given at $t = 0$.

Referring to Eq. (6–21) the solution $\mathbf{x}(t)$ can be written as follows:

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}_0 + \int_{0-}^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)\,d\tau$$

(1) *Impulse response:* Let us put the impulse input $\mathbf{u}(t)$ as

$$\mathbf{u}(t) = \delta(t)\,\mathbf{w}$$

where $\mathbf{w}$ is a vector whose components are the magnitudes of r impulse functions applied at $t = 0$. The solution to the impulse input $\delta(t)\,\mathbf{w}$ given at $t = 0$ is

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}_0 + \int_{0-}^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\,\delta(\tau)\,\mathbf{w}d\tau$$
$$= e^{\mathbf{A}t}\mathbf{x}_0 + e^{\mathbf{A}t}\mathbf{B}\mathbf{w}$$

(2) *Step response:* Let us put the step input $\mathbf{u}(t)$ as

$$\mathbf{u}(t) = \mathbf{k}$$

where $\mathbf{k}$ is a vector whose components are the magnitudes of r step functions applied at $t = 0$. The solution to the step input applied at $t = 0$ is given by

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}_0 + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{k}\,d\tau$$
$$= e^{\mathbf{A}t}\mathbf{x}_0 + \mathbf{A}^{-1}(e^{\mathbf{A}t} - \mathbf{I})\mathbf{B}\mathbf{k}$$

(3) *Ramp response:* Let us put the ramp input $\mathbf{u}(t)$ as

$$\mathbf{u}(t) = t\mathbf{v}$$

where $\mathbf{v}$ is a vector whose components are the magnitudes of ramp functions given at $t = 0$. The solution to the ramp input $t\mathbf{v}$ given at $t = 0$ is obtained as

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}_0 + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\tau\mathbf{v}\,d\tau$$
$$= e^{\mathbf{A}t}\mathbf{x}_0 + [\mathbf{A}^{-2}(e^{\mathbf{A}t} - \mathbf{I}) - \mathbf{A}^{-1}t]\,\mathbf{B}\mathbf{v}$$

6-4 METHODS FOR COMPUTING $e^{\mathbf{A}t}$ [1]

The solution of the linear time-invariant vector matrix differential equation involves the matrix exponential, $e^{\mathbf{A}t}$. It is necessary to express $e^{\mathbf{A}t}$ as a square matrix in order to obtain numerical solutions of $\mathbf{x}(t)$.

[1] This section follows (*6–4*, *6–5*). Figures in parentheses refer to references listed at the end of this chapter.

There are several methods available for computing $e^{\mathbf{A}t}$. If the number of rows of the square matrix becomes five or larger, hand computation becomes exceedingly tedious and machine computation may become necessary. (There are standard computational schemes of $e^{\mathbf{A}t}$ for machine computation.) We now present the following four methods:

Method 1 Transform **A** into the Jordan canonical form. If the eigenvectors of **A** are distinct, then referring to Eq. (6–19), we find that

$$e^{\mathbf{A}t} = \mathbf{P}e^{(\mathbf{P}^{-1}\mathbf{A}\mathbf{P})t}\mathbf{P}^{-1} = \mathbf{P}e^{\mathbf{D}t}\mathbf{P}^{-1}$$

(Refer also to Theorem 5–5.) Once **A** is given and the eigenvalues are determined, then a necessary transformation matrix **P** can be obtained by the methods presented in Chapter 3. $e^{\mathbf{A}t}$ can then be obtained as a square matrix as follows:

$$e^{\mathbf{A}t} = \mathbf{P}e^{\mathbf{D}t}\mathbf{P}^{-1} = \mathbf{P}\begin{bmatrix} e^{\lambda_1 t} & & & 0 \\ & e^{\lambda_2 t} & & \\ & & \ddots & \\ 0 & & & e^{\lambda_n t} \end{bmatrix}\mathbf{P}^{-1} \tag{6–25}$$

Equation (6–25) applies only to the case where **A** has distinct eigenvectors.

If **A** involves multiple eigenvectors, then $e^{\mathbf{A}t}$ can be written in a slightly complicated form. (Refer to Section 3–8.)

Method 2 Expand $e^{\mathbf{A}t}$ into a power series in t. (The matrix exponential $e^{\mathbf{A}t}$ is uniformly convergent for finite values of t.) Add the corresponding elements in the matrix terms of the infinite series. This method was discussed in Section 5–2. Except for very simple matrices **A** this method is not quite efficient for hand computation.

Since $e^{\mathbf{A}t}$ is given by

$$e^{\mathbf{A}t} = e^{\mathbf{M}} = \mathbf{I} + \mathbf{M} + \frac{\mathbf{M}}{2}\left(\frac{\mathbf{M}}{1!}\right) + \frac{\mathbf{M}}{3}\left(\frac{\mathbf{M}^2}{2!}\right) + \cdots + \frac{\mathbf{M}}{n}\left(\frac{\mathbf{M}^{n-1}}{(n-1)!}\right) + \cdots$$

note that each term in parentheses is equal to the entire preceding term. This provides a convenient recursion scheme. If a computer is available, this method may be used. The computation is carried out to only enough terms so that additional terms are negligible by comparison with the partial sum to that point.

In obtaining $e^{\mathbf{A}t}$ by the use of a digital computer the virtues of the series expansion technique are its simplicity and ease in programming. It is not necessary to find the eigenvalues of **A**. There is, however, some computational disadvantage to the series expansion method. This disadvantage stems from convergence requirements for the series $e^{\mathbf{A}t}$. The running time for a simulation will be the longest among other schemes for computing $e^{\mathbf{A}t}$. Use of the Jordan canonical form requires considerably more programming but will run in a fraction of time needed for the series solution.

The series expansion of $e^{\mathbf{A}t}$ fails to provide the insight into the system behavior afforded by the Jordan transformation, because the characteristic responses of different modes are not separated.

Method 3 Use the Laplace transform method. Taking the Laplace transform of the following equation:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}, \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{6–26}$$

we obtain

$$s\mathbf{X}(s) - \mathbf{x}_0 = \mathbf{A}\mathbf{X}(s)$$

or

$$(s\mathbf{I} - \mathbf{A})\mathbf{X}(s) = \mathbf{x}_0$$

from which

$$\mathbf{X}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{x}_0 \tag{6–27}$$

Since the solution of Eq. (6–26) is given by

$$\mathbf{x} = e^{\mathbf{A}t}\mathbf{x}_0$$

we obtain

$$\mathbf{X}(s) = \mathscr{L}[e^{\mathbf{A}t}]\mathbf{x}_0 \tag{6–28}$$

Comparing Eq. (6–27) and Eq. (6–28), we see that for $t \geqslant 0$

$$\mathscr{L}[e^{\mathbf{A}t}] = (s\mathbf{I} - \mathbf{A})^{-1}$$

Thus, to obtain $e^{\mathbf{A}t}$, first invert the matrix $(s\mathbf{I} - \mathbf{A})$. This results in a matrix whose elements are rational functions of s. Then, take the inverse Laplace transform of each element of the matrix. $(s\mathbf{I} - \mathbf{A})^{-1}$ can be obtained by computing the cofactors. This, however, may not be very efficient. In Section 6–5 we shall present two other methods for computing $(s\mathbf{I} - \mathbf{A})^{-1}$.

Method 4 Use Sylvester's interpolation formula. We shall first consider the case where the roots of the minimal polynomial $\phi(\lambda)$ of **A** are distinct. Then we shall deal with the case of multiple roots.

Case 1: where the minimal polynomial of **A** *involves only distinct roots* We shall assume that the degree of the minimal polynomial of **A** is m. By substituting $f(\mathbf{A}) = e^{\mathbf{A}t}$ into Eq. (5–8) we obtain

$$\begin{vmatrix} 1 & \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^{m-1} & e^{\lambda_1 t} \\ 1 & \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^{m-1} & e^{\lambda_2 t} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & \lambda_m & \lambda_m^2 & \cdots & \lambda_m^{m-1} & e^{\lambda_m t} \\ \mathbf{I} & \mathbf{A} & \mathbf{A}^2 & \cdots & \mathbf{A}^{m-1} & e^{\mathbf{A}t} \end{vmatrix} = 0 \tag{6–29}$$

By solving Eq. (6–29) for $e^{\mathbf{A}t}$, $e^{\mathbf{A}t}$ can be obtained in terms of the $\mathbf{A}^k$ ($k = 0, 1, 2, \ldots, m-1$) and the $e^{\lambda_i t}$ ($i = 1, 2, \ldots, m$). [Equation (6–29) may be expanded, for example, about the last column.]

Solving Eq. (6–29) for $e^{\mathbf{A}t}$ is the same as writing

$$e^{\mathbf{A}t} = \alpha_0\mathbf{I} + \alpha_1\mathbf{A} + \alpha_2\mathbf{A}^2 + \cdots + \alpha_{m-1}\mathbf{A}^{m-1}$$

and determining the α_k ($k = 0, 1, 2, \ldots, m-1$) by solving the following set of m equations for the α_k:

$$\begin{aligned}
\alpha_0 + \alpha_1\lambda_1 + \alpha_2\lambda_1^2 + \cdots + \alpha_{m-1}\lambda_1^{m-1} &= e^{\lambda_1 t}\\
\alpha_0 + \alpha_1\lambda_2 + \alpha_2\lambda_2^2 + \cdots + \alpha_{m-1}\lambda_2^{m-1} &= e^{\lambda_2 t}\\
\cdots \quad \cdots \quad \cdots &\\
\alpha_0 + \alpha_1\lambda_m + \alpha_2\lambda_m^2 + \cdots + \alpha_{m-1}\lambda_m^{m-1} &= e^{\lambda_m t}
\end{aligned}$$

If $\mathbf{A}$ is an $n \times n$ matrix and has distinct eigenvalues, then the number of the α_k to be determined is $m = n$. If $\mathbf{A}$ involves multiple eigenvalues but its minimal polynomial has only simple roots, however, then the number of the α_k to be determined is $m < n$.

Case 2: where the minimal polynomial of $\mathbf{A}$ *involves multiple roots* Suppose that the minimal polynomial of $\mathbf{A}$ involves three equal roots. Then, by substituting

$$f(\mathbf{A}) = e^{\mathbf{A}t}$$

into Eq. (5–18) we obtain

$$\begin{vmatrix}
0 & 0 & 1 & 3\lambda_1 & \cdots & \dfrac{(m-1)(m-2)}{2}\lambda_1^{m-3} & \dfrac{t^2}{2}e^{\lambda_1 t}\\
0 & 1 & 2\lambda_1 & 3\lambda_1^2 & \cdots & (m-1)\lambda_1^{m-2} & te^{\lambda_1 t}\\
1 & \lambda_1 & \lambda_1^2 & \lambda_1^3 & \cdots & \lambda_1^{m-1} & e^{\lambda_1 t}\\
1 & \lambda_4 & \lambda_4^2 & \lambda_4^3 & \cdots & \lambda_4^{m-1} & e^{\lambda_4 t}\\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots\\
1 & \lambda_m & \lambda_m^2 & \lambda_m^3 & \cdots & \lambda_m^{m-1} & e^{\lambda_m t}\\
\mathbf{I} & \mathbf{A} & \mathbf{A}^2 & \mathbf{A}^3 & \cdots & \mathbf{A}^{m-1} & e^{\mathbf{A}t}
\end{vmatrix} = 0 \qquad (6\text{–}30)$$

Equation (6–30) can be solved for $e^{\mathbf{A}t}$. [Equation (6–30) may be expanded, for example, about the last column.] Just like Case 1, solving Eq. (6–30) for $e^{\mathbf{A}t}$ is the same as writing $e^{\mathbf{A}t}$ as

$$e^{\mathbf{A}t} = \alpha_0\mathbf{I} + \alpha_1\mathbf{A} + \alpha_2\mathbf{A}^2 + \cdots + \alpha_{m-1}\mathbf{A}^{m-1}$$

and determining the α_k ($k = 0, 1, 2, \ldots, m-1$) from

$$\begin{aligned}
\alpha_2 + 3\alpha_3\lambda_1 + \cdots + \frac{(m-1)(m-2)}{2}\alpha_{m-1}\lambda_1^{m-3} &= \frac{t^2}{2}e^{\lambda_1 t}\\
\alpha_1 + 2\alpha_2\lambda_1 + 3\alpha_3\lambda_1^2 \cdots + (m-1)\alpha_{m-1}\lambda_1^{m-2} &= te^{\lambda_1 t}
\end{aligned}$$

$$\begin{aligned}
\alpha_0 + \alpha_1\lambda_1 + \alpha_2\lambda_1^2 + \cdots + \alpha_{m-1}\lambda_1^{m-1} &= e^{\lambda_1 t}\\
\alpha_0 + \alpha_1\lambda_4 + \alpha_2\lambda_4^2 + \cdots + \alpha_{m-1}\lambda_4^{m-1} &= e^{\lambda_4 t}\\
\cdots \quad \cdots \quad \cdots &\\
\alpha_0 + \alpha_1\lambda_m + \alpha_2\lambda_m^2 + \cdots + \alpha_{m-1}\lambda_m^{m-1} &= e^{\lambda_m t}
\end{aligned}$$

The extensions to other cases where, for example, there are two or more sets of multiple roots will be apparent. Note that if the minimal polynomial of $\mathbf{A}$ is not found, we may substitute the characteristic polynomial for the minimal polynomial. Amount of computations may, of course, be increased.

Example 6-5 Compute $e^{\mathbf{A}t}$ where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & -2 \end{bmatrix}$$

We shall illustrate computation of $e^{\mathbf{A}t}$ by use of the four methods presented earlier.

Method 1 The eigenvalues of $\mathbf{A}$ are 0 and -2. Hence, the necessary transformation matrix $\mathbf{P}$ is given as

$$\mathbf{P} = \begin{bmatrix} 1 & 1 \\ 0 & -2 \end{bmatrix}$$

Therefore from Eq. (6–25) $e^{\mathbf{A}t}$ is given as

$$\begin{aligned}
e^{\mathbf{A}t} &= \begin{bmatrix} 1 & 1 \\ 0 & -2 \end{bmatrix}\begin{bmatrix} e^0 & 0 \\ 0 & e^{-2t} \end{bmatrix}\begin{bmatrix} 1 & \frac{1}{2} \\ 0 & -\frac{1}{2} \end{bmatrix}\\
&= \begin{bmatrix} 1 & \frac{1}{2}(1 - e^{-2t}) \\ 0 & e^{-2t} \end{bmatrix}
\end{aligned}$$

Method 2 $e^{\mathbf{A}t}$ may be expanded as series of matrices and then added together into a closed form as follows:

$$\begin{aligned}
e^{\mathbf{A}t} &= \mathbf{I} + \sum_{k=1}^{\infty}\frac{\mathbf{A}^k t^k}{k!}\\
&= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 0 & -2 \end{bmatrix}t + \begin{bmatrix} 0 & 1 \\ 0 & -2 \end{bmatrix}^2\frac{t^2}{2!} + \cdots\\
&= \begin{bmatrix} 1 & \dfrac{1}{2} - \dfrac{1}{2}\left(1 - 2t + \dfrac{(2t)^2}{2!} - \dfrac{(2t)^3}{3!} + \cdots\right) \\ 0 & 1 - 2t + \dfrac{(2t)^2}{2!} - \dfrac{(2t)^3}{3!} + \dfrac{(2t)^4}{4!} - \cdots \end{bmatrix}\\
&= \begin{bmatrix} 1 & \frac{1}{2}(1 - e^{-2t}) \\ 0 & e^{-2t} \end{bmatrix}
\end{aligned}$$

Method 3 Since

$$s\mathbf{I} - \mathbf{A} = \begin{bmatrix} s & 0 \\ 0 & s \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 0 & -2 \end{bmatrix} = \begin{bmatrix} s & -1 \\ 0 & s+2 \end{bmatrix}$$

$(s\mathbf{I} - \mathbf{A})^{-1}$ is obtained as

$$(s\mathbf{I} - \mathbf{A})^{-1} = \begin{bmatrix} \frac{1}{s} & \frac{1}{s(s+2)} \\ 0 & \frac{1}{s+2} \end{bmatrix}$$

Then we have

$$\mathscr{L}[e^{\mathbf{A}t}] = (s\mathbf{I} - \mathbf{A})^{-1} = \begin{bmatrix} \frac{1}{s} & \frac{1}{s(s+2)} \\ 0 & \frac{1}{s+2} \end{bmatrix}$$

or

$$e^{\mathbf{A}t} = \begin{bmatrix} \mathscr{L}^{-1}\left[\frac{1}{s}\right] & \mathscr{L}^{-1}\left[\frac{1}{s(s+2)}\right] \\ \mathscr{L}^{-1}[0] & \mathscr{L}^{-1}\left[\frac{1}{s+2}\right] \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{2}(1 - e^{-2t}) \\ 0 & e^{-2t} \end{bmatrix}$$

Note that if $(s\mathbf{I} - \mathbf{A})$ is of complicated form, either of the two methods for computing $(s\mathbf{I} - \mathbf{A})^{-1}$ presented in Section 6–5 may be better in saving time for computation.

Method 4 Equation (6–29) becomes

$$\begin{vmatrix} 1 & \lambda_1 & e^{\lambda_1 t} \\ 1 & \lambda_2 & e^{\lambda_2 t} \\ \mathbf{I} & \mathbf{A} & e^{\mathbf{A}t} \end{vmatrix} = 0$$

Substituting $\lambda_1 = 0$, $\lambda_2 = -2$ into this last equation we obtain

$$\begin{vmatrix} 1 & 0 & 1 \\ 1 & -2 & e^{-2t} \\ \mathbf{I} & \mathbf{A} & e^{\mathbf{A}t} \end{vmatrix} = 0$$

Expanding the determinant, we obtain

$$-2\,e^{\mathbf{A}t} + \mathbf{A} + 2\mathbf{I} - \mathbf{A}e^{-2t} = \mathbf{0}$$

or

$$\begin{aligned} e^{\mathbf{A}t} &= \tfrac{1}{2}(\mathbf{A} + 2\mathbf{I} - \mathbf{A}e^{-2t}) \\ &= \tfrac{1}{2}\left\{\begin{bmatrix} 0 & 1 \\ 0 & -2 \end{bmatrix} + \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 0 & -2 \end{bmatrix} e^{-2t}\right\} \\ &= \begin{bmatrix} 1 & \frac{1}{2}(1 - e^{-2t}) \\ 0 & e^{-2t} \end{bmatrix} \end{aligned}$$

The alternate approach is to solve the following equations for α_0 and α_1:

$$\begin{aligned} \alpha_0 + \alpha_1\lambda_1 &= e^{\lambda_1 t} \\ \alpha_0 + \alpha_1\lambda_2 &= e^{\lambda_2 t} \end{aligned} \tag{6–31}$$

Since $\lambda_1 = 0$, $\lambda_2 = -2$, Eq. (6–31) becomes as follows:

$$\begin{aligned} \alpha_0 &= 1 \\ \alpha_0 - 2\alpha_1 &= e^{-2t} \end{aligned}$$

from which

$$\alpha_0 = 1, \qquad \alpha_1 = \tfrac{1}{2}(1 - e^{-2t})$$

Thus, $e^{\mathbf{A}t}$ can be written as

$$\begin{aligned} e^{\mathbf{A}t} &= \alpha_0\mathbf{I} + \alpha_1\mathbf{A} = \mathbf{I} + \tfrac{1}{2}(1 - e^{-2t})\mathbf{A} \\ &= \begin{bmatrix} 1 & \frac{1}{2}(1 - e^{-2t}) \\ 0 & e^{-2t} \end{bmatrix} \end{aligned}$$

Once $e^{\mathbf{A}t}$ is computed, the solution to the following vector matrix differential equation,

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

can be obtained as follows:

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}(0)$$

or

$$\begin{aligned} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} &= \begin{bmatrix} 1 & \frac{1}{2}(1 - e^{-2t}) \\ 0 & e^{-2t} \end{bmatrix} \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} \\ &= \begin{bmatrix} x_1(0) + \frac{1}{2}x_2(0) - \frac{1}{2}e^{-2t}x_2(0) \\ e^{-2t}x_2(0) \end{bmatrix} \end{aligned}$$

Example 6-6 Matrix Riccati equation (6–6) Consider the matrix differential equation

$$\dot{\mathbf{X}} = \mathbf{A}(t)\mathbf{X} + \mathbf{X}\mathbf{B}(t) + \mathbf{X}\mathbf{C}(t)\mathbf{X} + \mathbf{D}(t)$$

where $\mathbf{X}$ is an $n \times n$ matrix and $\mathbf{A}(t)$, $\mathbf{B}(t)$, $\mathbf{C}(t)$, and $\mathbf{D}(t)$ are also $n \times n$ matrices. It is assumed that $\mathbf{A}(t)$, $\mathbf{B}(t)$, $\mathbf{C}(t)$, and $\mathbf{D}(t)$ are absolutely integrable as functions of t in the interval $t_0 \leqslant t \leqslant t_1$. This equation is called the *matrix Riccati equation.* The matrix Riccati equation arises frequently in the solution of optimal control problems. We shall show that the matrix Riccati equation can be solved by using two simultaneous linear matrix differential equations.

Let us define two simultaneous matrix differential equations,

$$\begin{aligned} \dot{\mathbf{Y}} &= -\mathbf{B}(t)\mathbf{Y} - \mathbf{C}(t)\mathbf{Z} \\ \dot{\mathbf{Z}} &= \mathbf{D}(t)\mathbf{Y} + \mathbf{A}(t)\mathbf{Z} \end{aligned} \tag{6–32}$$

If in the interval $t_0 \leqslant t \leqslant t_1$ the $n \times n$ matrix $\mathbf{Y}$ is found to be nonsingular, then the solution to the matrix Riccati equation can be given by

$$\mathbf{X}(t) = \mathbf{Z}(t)\mathbf{Y}^{-1}(t) \qquad t_0 \leqslant t \leqslant t_1$$

This can be easily verified. Since

$$\dot{\mathbf{X}} = \dot{\mathbf{Z}}\mathbf{Y}^{-1} + \mathbf{Z}\dot{\mathbf{Y}}^{-1} \qquad \text{and} \qquad \dot{\mathbf{Y}}^{-1} = -\mathbf{Y}^{-1}\dot{\mathbf{Y}}\mathbf{Y}^{-1}$$

we obtain

$$\begin{aligned} \dot{\mathbf{X}} &= [\mathbf{D}(t)\mathbf{Y} + \mathbf{A}(t)\mathbf{Z}]\mathbf{Y}^{-1} + \mathbf{Z}(-\mathbf{Y}^{-1})[-\mathbf{B}(t)\mathbf{Y} - \mathbf{C}(t)\mathbf{Z}]\mathbf{Y}^{-1} \\ &= \mathbf{D}(t) + \mathbf{A}(t)\mathbf{X} + \mathbf{X}\mathbf{B}(t) + \mathbf{X}\mathbf{C}(t)\mathbf{X} \end{aligned}$$

If $\mathbf{D}(t)$ is zero, then $\mathbf{Z}$ may be solved independently of $\mathbf{Y}$. Once $\mathbf{Z}$ is obtained, $\mathbf{Y}$ may be solved. In general, however, if coefficient matrices in the matrix

Riccati equation are time-varying, solving Eq. (6–32) for $\mathbf{Y}$ and $\mathbf{Z}$ is not a simple matter. If, however, $\mathbf{A}(t)$, $\mathbf{B}(t)$, $\mathbf{C}(t)$, and $\mathbf{D}(t)$ are constant matrices, then it is easy to obtain at least a formal solution for $\mathbf{Y}(t)$ and $\mathbf{Z}(t)$.

If matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are constant matrices, choose $t_0 = 0$ and define a $2n \times n$ matrix $\mathbf{W}(t)$ and $2n \times 2n$ matrix $\mathbf{G}$ as follows:

$$\mathbf{W}(t) = \begin{bmatrix} \mathbf{Y}(t) \\ \hline \mathbf{Z}(t) \end{bmatrix}, \qquad \mathbf{G} = \left[\begin{array}{c|c} -\mathbf{B} & -\mathbf{C} \\ \hline \mathbf{D} & \mathbf{A} \end{array}\right]$$

Then Eq. (6–32) can be represented by

$$\dot{\mathbf{W}}(t) = \mathbf{G}\mathbf{W}(t)$$

whose solution is

$$\mathbf{W}(t) = e^{\mathbf{G}t}\mathbf{W}(0)$$

Once $e^{\mathbf{G}t}$ can be obtained as a square matrix, then $\mathbf{Y}(t)$ and $\mathbf{Z}(t)$ can be determined immediately. If $\mathbf{Y}(t)$ is nonsingular, then the solution $\mathbf{X}(t)$ to the time-invariant matrix Riccati equation is given by

$$\mathbf{X}(t) = \mathbf{Z}(t)\mathbf{Y}^{-1}(t)$$

6-5 TRANSFER MATRICES

In this section, we present s plane solution of the linear time-invariant vector matrix differential equation. We then define the transfer matrix. Finally, we present two methods for computing $(s\mathbf{I} - \mathbf{A})^{-1}$.

Consider the system described by

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \tag{6–33}$$

where

$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{A} = n \times n$ constant matrix
$\mathbf{B} = n \times r$ constant matrix

The Laplace transform of Eq. (6–33) with zero initial condition is

$$s\mathbf{X}(s) = \mathbf{A}\mathbf{X}(s) + \mathbf{B}\mathbf{U}(s)$$

from which

$$\mathbf{X}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{U}(s) \tag{6–34}$$

The poles of this equation are the values of s for which the determinant is zero, namely,

$$|s\mathbf{I} - \mathbf{A}| = 0$$

This is the equation determining the eigenvalues of the matrix $\mathbf{A}$. Hence the poles of the system are the eigenvalues of $\mathbf{A}$.

Transfer matrix Consider the system shown in Fig. 6–3. From the block diagram

$$\begin{aligned} X_1(s) &= H_{11}(s)U_1(s) + H_{12}(s)U_2(s) \\ X_2(s) &= H_{21}(s)U_1(s) + H_{22}(s)U_2(s) \end{aligned} \tag{6–35}$$

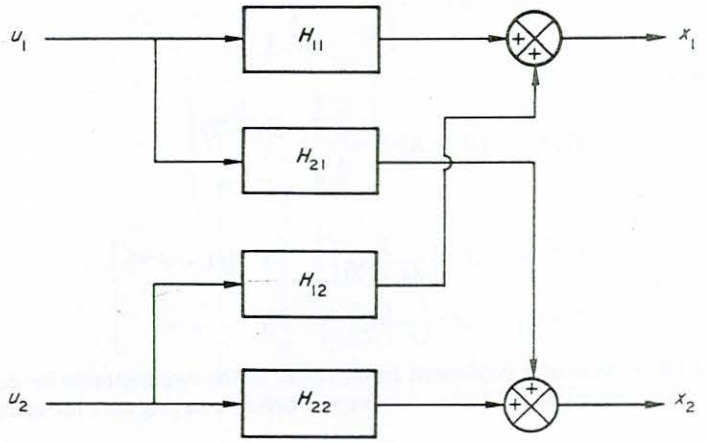


Figure 6-3 Multiple-input multiple-output system.

In such a multiple-input multiple-output system, the transfer functions constitute a transfer matrix. Rewriting Eq. (6–35) in terms of vector matrix notation

$$\begin{bmatrix} X_1(s) \\ X_2(s) \end{bmatrix} = \begin{bmatrix} H_{11}(s) & H_{12}(s) \\ H_{21}(s) & H_{22}(s) \end{bmatrix} \begin{bmatrix} U_1(s) \\ U_2(s) \end{bmatrix}$$

In this system the transfer matrix consists of four components H_{11}, H_{12}, H_{21}, and H_{22}. In the transfer matrix the (i, j)th element is the transfer function between the ith output and jth input. It is the Laplace transform of the unit impulse response, the impulse being applied at the jth input and response being taken at the ith output.

A state space equation for such a system defined in terms of the transfer matrix can be obtained easily if the transfer matrix $\mathbf{H}(s)$ relating the state vector $\mathbf{x}$ and the control vector $\mathbf{u}$ is a square matrix and is nonsingular.

As an example consider the following system

$$\begin{bmatrix} X_1(s) \\ X_2(s) \end{bmatrix} = \begin{bmatrix} \dfrac{1}{s} & \dfrac{1}{s(s+2)} \\ 0 & \dfrac{1}{s+2} \end{bmatrix} \begin{bmatrix} U_1(s) \\ U_2(s) \end{bmatrix} \tag{6–36}$$

Let us find a state space equation of this system in the following standard form:

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}$$

From Eq. (6–34) and Eq. (6–36)

$$(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} = \begin{bmatrix} \frac{1}{s} & \frac{1}{s(s+2)} \\ 0 & \frac{1}{s+2} \end{bmatrix}$$

Hence

$$\mathbf{B}^{-1}(s\mathbf{I} - \mathbf{A}) = \begin{bmatrix} \frac{1}{s} & \frac{1}{s(s+2)} \\ 0 & \frac{1}{s+2} \end{bmatrix}^{-1} = \begin{bmatrix} s & -1 \\ 0 & s+2 \end{bmatrix}$$

$$= s\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 0 & -2 \end{bmatrix}$$

Comparing the coefficients of the s terms and constant terms on both sides of this last equation we obtain

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & -2 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Hence the state space equation becomes

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

The eigenvalues of $\mathbf{A}$ are 0 and -2. These eigenvalues are clearly recognized in the transfer matrix.

One way to obtain a state space equation of the system given by $\mathbf{X}(s) = \mathbf{H}(s)\,\mathbf{U}(s)$ in which the transfer matrix $\mathbf{H}(s)$ is not a nonsingular matrix is to draw an analog computer diagram for simulation of the system. Then the output signals from the integrators are a set of state variables for the system. The state space equation for such a case can be written directly from the analog computer diagram.

Computation of $(s\mathbf{I} - \mathbf{A})^{-1}$ If we wish to determine the transfer matrix from Eq. (6–33), we must compute $(s\mathbf{I} - \mathbf{A})^{-1}$. Also, in computing $e^{\mathbf{A}t}$ by the Laplace transform method we need computing $(s\mathbf{I} - \mathbf{A})^{-1}$. Computing $(s\mathbf{I} - \mathbf{A})^{-1}$ is, except in very simple cases, generally a time-consuming task. The simplest but time-consuming way to obtain the $n \times n$ matrix $(s\mathbf{I} - \mathbf{A})^{-1}$ is to compute cofactors. This is not efficient if $n > 3$. There are a few other methods for computing $(s\mathbf{I} - \mathbf{A})^{-1}$. We present two such methods in what follows:

Method 1[1] Let us put

$$\mathbf{P} = (s\mathbf{I} - \mathbf{A})^{-1}$$

then

$$s\mathbf{P} = \mathbf{AP} + \mathbf{I}$$

By premultiplying $(s\mathbf{I} + \mathbf{A})$ to both sides of this equation we obtain

$$s^2\mathbf{P} = \mathbf{A}^2\mathbf{P} + \mathbf{A} + s\mathbf{I}$$

Similarly, by premultiplying $(s\mathbf{I} + \mathbf{A})$ to both sides of this last equation, we obtain

$$s^3\mathbf{P} = \mathbf{A}^3\mathbf{P} + \mathbf{A}^2 + s\mathbf{A} + s^2\mathbf{I}$$

By repeating this process we obtain the following set of equations

$$\begin{aligned} \mathbf{P} &= \mathbf{P} \\ s\mathbf{P} &= \mathbf{AP} + \mathbf{I} \\ s^2\mathbf{P} &= \mathbf{A}^2\mathbf{P} + \mathbf{A} + s\mathbf{I} \\ s^3\mathbf{P} &= \mathbf{A}^3\mathbf{P} + \mathbf{A}^2 + s\mathbf{A} + s^2\mathbf{I} \\ &\cdots \qquad \cdots \qquad \cdots \\ s^p\mathbf{P} &= \mathbf{A}^p\mathbf{P} + \mathbf{A}^{p-1} + s\mathbf{A}^{p-2} + \cdots + s^{p-2}\mathbf{A} + s^{p-1}\mathbf{I} \end{aligned}$$

where p is the degree of the minimal polynomial of $\mathbf{A}$. Let the minimal polynomial of $\mathbf{A}$ be

$$\alpha_0\mathbf{I} + \alpha_1\mathbf{A} + \alpha_2\mathbf{A}^2 + \cdots + \alpha_p\mathbf{A}^p = \mathbf{0}$$

where $\alpha_p = 1$. Then

$$\begin{aligned} \sum_{i=0}^{p} \alpha_i s^i\mathbf{P} &= \sum_{i=0}^{p} \alpha_i\mathbf{A}^i\mathbf{P} + \sum_{i=1}^{p} \alpha_i\mathbf{A}^{i-1} + s\sum_{i=2}^{p} \alpha_i\mathbf{A}^{i-2} \\ &\quad + \cdots + s^{p-2}\sum_{i=p-1}^{p} \alpha_i\mathbf{A}^{i-p+1} + s^{p-1}\alpha_p\mathbf{I} \\ &= \sum_{j=0}^{p-1} s^j \sum_{i=1+j}^{p} \alpha_i\mathbf{A}^{i-j-1} \end{aligned}$$

Therefore,

$$(s\mathbf{I} - \mathbf{A})^{-1} = \mathbf{P} = \frac{\sum_{j=0}^{p-1} s^j \sum_{i=1+j}^{p} \alpha_i\mathbf{A}^{i-j-1}}{\sum_{i=0}^{p} \alpha_i s^i} \tag{6–37}$$

If the minimal polynomial and characteristic polynomial of $\mathbf{A}$ are identical, then $p = n$. If $p = n$, then Eq. (6–37) becomes

$$(s\mathbf{I} - \mathbf{A})^{-1} = \frac{\sum_{j=0}^{n-1} s^j \sum_{i=1+j}^{n} \alpha_i\mathbf{A}^{i-j-1}}{|s\mathbf{I} - \mathbf{A}|} \tag{6–38}$$

where

$$|s\mathbf{I} - \mathbf{A}| = \sum_{i=0}^{n} \alpha_i s^i, \qquad \alpha_n = 1$$

Method 2 (6–4) The second method is based on Sylvester's interpolation formula given by Eq. (5–6). We first consider the case where the minimal polynomial of $\mathbf{A}$ has only distinct roots and then consider the case of multiple roots.

[1] Discussion of Method 1 follows closely (6–7).

Case 1: where the minimal polynomial of **A** *involves only distinct roots* Let us put

$$f(\mathbf{A}) = (s\mathbf{I} - \mathbf{A})^{-1}$$

The function $f(\mathbf{A})$ can be written as

$$f(\mathbf{A}) = \sum_{k=1}^{p} f(s_k)\mathbf{Z}_k \tag{6–39}$$

where p is the degree of the minimal polynomial of **A**, and the $\mathbf{Z}_k$ are $n \times n$ matrices. Equation (6–39) can be written as

$$(s\mathbf{I} - \mathbf{A})^{-1} = \sum_{k=1}^{p} \frac{\mathbf{Z}_k}{s - s_k} \tag{6–40}$$

To determine the $\mathbf{Z}_k$, we substitute

$$g_i(s) = (s - s_k)^{i-1} \qquad (i = 1, 2, \ldots, p)$$

where s_k is any one of the p roots of the minimal polynomial of **A**, into the following p simultaneous equations and solve for the $\mathbf{Z}_k$:

$$g_i(\mathbf{A}) = g_i(s_1)\mathbf{Z}_1 + g_i(s_2)\mathbf{Z}_2 + \cdots + g_i(s_p)\mathbf{Z}_p \qquad (i = 1, 2, \ldots, p) \tag{6–41}$$

Note that $g_i(\mathbf{A}) = (\mathbf{A} - s_k\mathbf{I})^{i-1}$. The p simultaneous equations possess a solution, since

$$\begin{vmatrix} g_1(s_1) & g_1(s_2) & \cdots & g_1(s_p) \\ g_2(s_1) & g_2(s_2) & \cdots & g_2(s_p) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ g_p(s_1) & g_p(s_2) & \cdots & g_p(s_p) \end{vmatrix}$$

$$= \begin{vmatrix} 1 & 1 & \cdots & 1 \\ s_1 - s_k & s_2 - s_k & \cdots & s_p - s_k \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ (s_1 - s_k)^{p-1} & (s_2 - s_k)^{p-1} & \cdots & (s_p - s_k)^{p-1} \end{vmatrix} \neq 0$$

As an example, consider the following 2×2 matrix **A**:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & -2 \end{bmatrix}$$

The eigenvalues of **A** are $s_1 = 0$ and $s_2 = -2$. Hence,

$$(s\mathbf{I} - \mathbf{A})^{-1} = \frac{\mathbf{Z}_1}{s} + \frac{\mathbf{Z}_2}{s + 2}$$

To determine $\mathbf{Z}_1$ and $\mathbf{Z}_2$, put

$$g_i(s) = (s - s_2)^{i-1} \qquad (i = 1, 2)$$

(We can, of course, take s_1 as s_k.) Then Eq. (6–41) becomes as follows:

$$g_1(\mathbf{A}) = g_1(s_1)\mathbf{Z}_1 + g_1(s_2)\mathbf{Z}_2$$
$$g_2(\mathbf{A}) = g_2(s_1)\mathbf{Z}_1 + g_2(s_2)\mathbf{Z}_2$$

or

$$\mathbf{I} = \mathbf{Z}_1 + \mathbf{Z}_2$$
$$\mathbf{A} + 2\mathbf{I} = 2\mathbf{Z}_1$$

Hence,

$$\mathbf{Z}_1 = \mathbf{I} + \tfrac{1}{2}\mathbf{A} = \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & 0 \end{bmatrix}$$

$$\mathbf{Z}_2 = \mathbf{I} - \mathbf{Z}_1 = -\tfrac{1}{2}\mathbf{A} = \begin{bmatrix} 0 & -\frac{1}{2} \\ 0 & 1 \end{bmatrix}$$

it follows that

$$(s\mathbf{I} - \mathbf{A})^{-1} = \frac{\mathbf{Z}_1}{s} + \frac{\mathbf{Z}_2}{s + 2} = \begin{bmatrix} \frac{1}{s} & \frac{1}{s(s+2)} \\ 0 & \frac{1}{s+2} \end{bmatrix}$$

Case 2: where the minimal polynomial of **A** *involves multiple roots* In this case $f(\mathbf{A})$ can be written as

$$f(\mathbf{A}) = \sum_{k=1}^{q} [f(s_k)\mathbf{Z}_{k1} + f'(s_k)\mathbf{Z}_{k2} + \cdots + f^{(m_k-1)}(s_k)\mathbf{Z}_{km_k}]$$

where m_k = the multiplicity of the factor $(s - s_k)$ in the minimal polynomial of **A**
q = the number of distinct roots $s_1, s_2, \ldots, s_q$ $(q < p)$ of the minimal polynomial of **A**
$p = m_1 + m_2 + \cdots + m_q$ = the degree of the minimal polynomial of **A**

In the present case

$$f(s_k) = (s - s_k)^{-1}$$
$$f'(s_k) = (s - s_k)^{-2}$$
$$\cdots \quad \cdots \quad \cdots$$
$$f^{(m_k-1)}(s_k) = (m_k - 1)!(s - s_k)^{-m_k}$$

where the derivatives are with respect to s_k. Hence we have

$$(s\mathbf{I} - \mathbf{A})^{-1} = \sum_{k=1}^{q} \left[\frac{\mathbf{Z}_{k1}}{s - s_k} + \frac{\mathbf{Z}_{k2}}{(s - s_k)^2} + \cdots + \frac{(m_k - 1)!\mathbf{Z}_{km_k}}{(s - s_k)^{m_k}}\right] \tag{6–42}$$

To determine the $\mathbf{Z}_{kj}$, we substitute

$$g_i(s) = (s - s_k)^{i-1} \qquad (i = 1, 2, \ldots, p)$$

where s_k is any one of the roots of the minimal polynomial of **A**, into the following p simultaneous equations and solve for the $\mathbf{Z}_{kj}$.

$$g_i(\mathbf{A}) = g_i(s_1)\mathbf{Z}_{11} + g_i'(s_1)\mathbf{Z}_{12} + \cdots + g_i^{(m_1-1)}(s_1)\mathbf{Z}_{1m_1}$$
$$+ g_i(s_2)\mathbf{Z}_{21} + \cdots + g_i^{(m_q-1)}(s_q)\mathbf{Z}_{qm_q} \tag{6–43}$$
$$(i = 1, 2, \ldots, p)$$

where the derivatives are with respect to s. Note that a solution to Eq. (6–43) exists, since the following matrix of coefficients is nonsingular:

$$\begin{bmatrix} g_1(s_1) & g_1'(s_1) & \cdots & g_1^{(m_1-1)}(s_1) & g_1(s_2) & \cdots & g_1^{(m_q-1)}(s_q) \\ g_2(s_1) & g_2'(s_1) & \cdots & g_2^{(m_1-1)}(s_1) & g_2(s_2) & \cdots & g_2^{(m_q-1)}(s_q) \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ g_p(s_1) & g_p'(s_1) & \cdots & g_p^{(m_1-1)}(s_1) & g_p(s_2) & \cdots & g_p^{(m_q-1)}(s_q) \end{bmatrix}$$

Example 6-7 (6-2) Consider a system described by

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}$$
$$\mathbf{y} = \mathbf{Cx} + \mathbf{Du} \tag{6-44}$$

where

$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{y} = m$ vector (output vector)
$\mathbf{A} = n \times n$ constant matrix
$\mathbf{B} = n \times r$ constant matrix
$\mathbf{C} = m \times n$ constant matrix
$\mathbf{D} = m \times r$ constant matrix

Let us determine the transfer matrices for this system. The Laplace transforms of Eq. (6-44) with the zero initial condition become

$$s\mathbf{X}(s) = \mathbf{AX}(s) + \mathbf{BU}(s)$$
$$\mathbf{Y}(s) = \mathbf{CX}(s) + \mathbf{DU}(s)$$

from which

$$\mathbf{X}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{BU}(s)$$
$$\mathbf{Y}(s) = [\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}]\mathbf{U}(s)$$

The transfer matrix between $\mathbf{X}(s)$ and $\mathbf{U}(s)$ is $(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$ and that between $\mathbf{Y}(s)$ and $\mathbf{U}(s)$ is $[\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}]$. The partial-fraction expansion of $\mathbf{X}(s)$ can be obtained by use of Eq. (6-42). The partial-fraction expansion of $\mathbf{Y}(s)$ is given by

$$\mathbf{Y}(s) = \left\{\mathbf{C}\sum_{k=1}^{q}\left[\frac{\mathbf{Z}_{k1}}{s - s_k} + \frac{\mathbf{Z}_{k2}}{(s - s_k)^2} + \cdots + \frac{(m_k - 1)!\mathbf{Z}_{km_k}}{(s - s_k)^{m_k}}\right]\mathbf{B} + \mathbf{D}\right\}\mathbf{U}(s) \tag{6-45}$$

Let us compute the transfer matrix between $\mathbf{X}(s)$ and $\mathbf{U}(s)$ and that between $Y(s)$ and $\mathbf{U}(s)$ for the following $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$:

$$\mathbf{A} = \begin{bmatrix} -5 & -4 & 2 \\ 3 & 3 & -2 \\ 0 & 2 & -2 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} -1 & 0 \\ 1 & 0 \\ 0 & 2 \end{bmatrix}$$

$$\mathbf{C} = [1 \quad 1 \quad 0], \qquad \mathbf{D} = [0 \quad 0]$$

In this system the roots of the minimal polynomial of $\mathbf{A}$ are $-1, -1, -2$.

To obtain the $\mathbf{Z}_{kj}$ put

$$g_1(s) = 1, \qquad g_2(s) = s + 1, \qquad g_3(s) = (s + 1)^2$$

Then we obtain the following three equations for the determination of the $\mathbf{Z}_{kj}$:

$$\begin{aligned} g_1(\mathbf{A}) &= g_1(s_1)\mathbf{Z}_{11} + g_1'(s_1)\mathbf{Z}_{12} + g_1(s_2)\mathbf{Z}_{21} \\ g_2(\mathbf{A}) &= g_2(s_1)\mathbf{Z}_{11} + g_2'(s_1)\mathbf{Z}_{12} + g_2(s_2)\mathbf{Z}_{21} \\ g_3(\mathbf{A}) &= g_3(s_1)\mathbf{Z}_{11} + g_3'(s_1)\mathbf{Z}_{12} + g_3(s_2)\mathbf{Z}_{21} \end{aligned} \tag{6-46}$$

Substituting

$$g_1(\mathbf{A}) = \mathbf{I}, \qquad g_2(\mathbf{A}) = \mathbf{A} + \mathbf{I}, \qquad g_3(\mathbf{A}) = (\mathbf{A} + \mathbf{I})^2$$

and

$$s_1 = -1, \qquad s_2 = -2$$

into Eq. (6-46) we obtain

$$\mathbf{I} = \mathbf{Z}_{11} + \mathbf{Z}_{21}$$
$$\mathbf{A} + \mathbf{I} = \mathbf{Z}_{12} - \mathbf{Z}_{21}$$
$$(\mathbf{A} + \mathbf{I})^2 = \mathbf{Z}_{21}$$

or

$$\mathbf{Z}_{11} = \begin{bmatrix} -3 & -4 & 2 \\ 0 & 1 & 0 \\ -6 & -6 & 4 \end{bmatrix}, \qquad \mathbf{Z}_{12} = \begin{bmatrix} 0 & 0 & 0 \\ 3 & 4 & -2 \\ 6 & 8 & -4 \end{bmatrix}, \qquad \mathbf{Z}_{21} = \begin{bmatrix} 4 & 4 & -2 \\ 0 & 0 & 0 \\ 6 & 6 & -3 \end{bmatrix}$$

$(s\mathbf{I} - \mathbf{A})^{-1}$ is then obtained as

$$(s\mathbf{I} - \mathbf{A})^{-1} = \frac{\mathbf{Z}_{11}}{s - s_1} + \frac{\mathbf{Z}_{12}}{(s - s_1)^2} + \frac{\mathbf{Z}_{21}}{s - s_2}$$

and $\mathbf{X}(s)$ is computed as

$$\begin{aligned} \mathbf{X}(s) &= (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{BU}(s) \\ &= \begin{bmatrix} \dfrac{-1}{s+1} & \dfrac{4}{(s+1)(s+2)} \\ \dfrac{s+2}{(s+1)^2} & \dfrac{-4}{(s+1)^2} \\ \dfrac{2}{(s+1)^2} & \dfrac{2(s-1)(s+3)}{(s+1)^2(s+2)} \end{bmatrix} \begin{bmatrix} U_1(s) \\ U_2(s) \end{bmatrix} \end{aligned} \tag{6-47}$$

The transfer matrix between $\mathbf{X}(s)$ and $\mathbf{U}(s)$ is given by this 3×2 matrix. From Eq. (6-47), $x_1(t)$, $x_2(t)$, and $x_3(t)$ for any inputs u_1 and u_2 can be computed easily.

The Laplace transform of the over-all response from the two inputs to the single output is given by

$$\begin{aligned} Y(s) &= [\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}]\mathbf{U}(s) \\ &= \begin{bmatrix} \dfrac{1}{(s+1)^2} & \dfrac{-4}{(s+1)^2(s+2)} \end{bmatrix} \begin{bmatrix} U_1(s) \\ U_2(s) \end{bmatrix} \end{aligned} \tag{6-48}$$

The output response for any known inputs u_1 and u_2 can be easily obtained from Eq. (6-48).

6-6 IMPULSE RESPONSE FUNCTIONS OF DYNAMIC SYSTEMS

In this section we obtain impulse response functions of dynamic systems and find responses of such systems to arbitrary inputs in terms of impulse response functions.

Consider first the following linear time-invariant differential equation:

$$x^{(n)} + a_1 x^{(n-1)} + a_2 x^{(n-2)} + \cdots + a_{n-1}\dot{x} + a_n x = u \tag{6-49}$$

where the a_i ($i = 1, 2, \ldots, n$) are constant. Equation (6–49) can be written as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \cdot \\ \cdot \\ \cdot \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix} [u]$$

where

$$\begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} x \\ \dot{x} \\ \cdot \\ \cdot \\ \cdot \\ x^{(n-1)} \end{bmatrix}$$

Let us define

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_1 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix}$$

With the initial time $t_0 = 0$, the solution of this equation can be obtained in the following form:

$$\mathbf{x}(t) = \Phi(t)\mathbf{x}(0) + \int_0^t \Phi(t-\tau)\mathbf{B}(\tau)u(\tau)\,d\tau$$

where

$$\begin{aligned} \Phi(t) &= [\Phi_1(t) \mid \Phi_2(t) \mid \cdots \mid \Phi_n(t)] \\ &= \begin{bmatrix} \phi_{11}(t) & \phi_{12}(t) & \cdots & \phi_{1n}(t) \\ \phi_{21}(t) & \phi_{22}(t) & \cdots & \phi_{2n}(t) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \phi_{n1}(t) & \phi_{n2}(t) & \cdots & \phi_{nn}(t) \end{bmatrix} \\ &= e^{\mathbf{A}t} \end{aligned}$$

[$\Phi_1(t), \Phi_2(t), \ldots, \Phi_n(t)$ are linearly independent.] Notice that if we define the eigenvalues of $\mathbf{A}$ as $\lambda_1, \lambda_2, \ldots, \lambda_n$, not necessarily distinct, then the ϕ_{ij} ($i, j = 1, 2, \ldots, n$) are linear combinations of the terms like $t^h e^{\lambda_i t}$ ($h = 0, 1, 2, \ldots, m_i - 1$, where m_i is the multiplicity of the eigenvalue λ_i.) Hence for $t > 0$ the ϕ_{ij} may be differentiated with respect to t as many times as we need.

In terms of the $\phi_{ij}(t)$, $x_1(t)$ is obtained as

$$\begin{aligned} x_1(t) = \phi_{11}(t)x_1(0) + \phi_{12}(t)x_2(0) + \cdots + \phi_{1n}(t)x_n(0) \\ + \int_0^t \phi_{1n}(t-\tau)u(\tau)\,d\tau \end{aligned} \tag{6–50}$$

If the initial conditions are zero, or $x_1(0) = x_2(0) = \cdots = x_n(0) = 0$, then

$$x_1(t) = \int_0^t \phi_{1n}(t-\tau)u(\tau)\,d\tau \tag{6–51}$$

Notice that $\phi_{1n}(t-\tau)$ is the impulse response function of Eq. (6–49). To verify this statement let us write the transfer function of the system given by Eq. (6–49).

$$\frac{X_1(s)}{U(s)} = \frac{1}{s^n + a_1 s^{n-1} + \cdots + a_{n-1}s + a_n}$$

Let us put

$$H(s) = \frac{1}{s^n + a_1 s^{n-1} + \cdots + a_{n-1}s + a_n}$$

Then

$$X_1(s) = H(s)U(s) \tag{6–52}$$

The inverse Laplace transform of Eq. (6–52) yields

$$x_1(t) = \int_0^t h(t-\tau)u(\tau)\,d\tau \tag{6–53}$$

where $h(t-\tau) =$ impulse response function

Comparing Eq. (6–51) and Eq. (6–53) we see that

$$\phi_{1n}(t-\tau) = h(t-\tau) \tag{6–54}$$

Notice that

$$\Phi_n(t) = \begin{bmatrix} \phi_{1n}(t) \\ \phi_{2n}(t) \\ \cdot \\ \cdot \\ \cdot \\ \phi_{nn}(t) \end{bmatrix} = \begin{bmatrix} h(t) \\ \dot{h}(t) \\ \cdot \\ \cdot \\ \cdot \\ h^{(n-1)}(t) \end{bmatrix}$$

Referring to Eq. (6–50) and Eq. (6–54) the solution of Eq. (6–49) can be written as follows:

$$\begin{aligned} x(t) = \phi_{11}(t)x(0) + \phi_{12}(t)\dot{x}(0) + \cdots + \phi_{1n}(t)x^{(n-1)}(0) \\ + \int_0^t h(t-\tau)u(\tau)\,d\tau \end{aligned}$$

We can easily extend the analysis made for the linear time-invariant system to the following linear time-varying system:

$$\overset{(n)}{x} + a_1(t)\overset{(n-1)}{x} + \cdots + a_{n-1}(t)\dot{x} + a_n(t)x = u \qquad (6\text{–}55)$$

where the $a_i(t)$ are absolutely integrable as functions of t in the interval $t_0 \leqslant t \leqslant t_1$. The response $x(t)$ depends on both the time of application of the input, t_0, and the time of measurement, t. The response of this system can be given by

$$x(t) = \phi_{11}(t, t_0)x(t_0) + \phi_{12}(t, t_0)\dot{x}(t_0) + \cdots + \phi_{1n}(t, t_0)\overset{(n-1)}{x}(t_0) + \int_{t_0}^{t} h(t, \tau)u(\tau)\, d\tau$$

where

$$h(t, \tau) = \phi_{11}(t, t_0)\psi_{1n}(\tau, t_0) + \phi_{12}(t, t_0)\psi_{2n}(\tau, t_0) + \phi_{13}(t, t_0)\psi_{3n}(\tau, t_0) + \cdots + \phi_{1n}(t, t_0)\psi_{nn}(\tau, t_0)$$

Here the $\phi_{1i}(t, t_0)$ and $\psi_{in}(\tau, t_0)$ are defined by

$$\Phi(t, t_0) = \begin{bmatrix} \phi_{11}(t, t_0) & \phi_{12}(t, t_0) & \cdots & \phi_{1n}(t, t_0) \\ \phi_{21}(t, t_0) & \phi_{22}(t, t_0) & \cdots & \phi_{2n}(t, t_0) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \phi_{n1}(t, t_0) & \phi_{n2}(t, t_0) & \cdots & \phi_{nn}(t, t_0) \end{bmatrix}$$

and

$$\Phi^{-1}(\tau, t_0) = \begin{bmatrix} \psi_{11}(\tau, t_0) & \psi_{12}(\tau, t_0) & \cdots & \psi_{1n}(\tau, t_0) \\ \psi_{21}(\tau, t_0) & \psi_{22}(\tau, t_0) & \cdots & \psi_{2n}(\tau, t_0) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \psi_{n1}(\tau, t_0) & \psi_{n2}(\tau, t_0) & \cdots & \psi_{nn}(\tau, t_0) \end{bmatrix}$$

If the initial conditions are zero, or $x(t_0) = \dot{x}(t_0) = \cdots = \overset{(n-1)}{x}(t_0) = 0$, then the response of the system of Eq. (6–55) to any $u(t)$ can be obtained from

$$x(t) = \int_{t_0}^{t} h(t, \tau)u(\tau)\, d\tau$$

The impulse response function $h(t, \tau)$ satisfies the following equation:

$$\frac{\partial^n h(t, \tau)}{\partial t^n} + a_1(t)\frac{\partial^{n-1} h(t, \tau)}{\partial t^{n-1}} + \cdots + a_{n-1}(t)\frac{\partial h(t, \tau)}{\partial t} + a_n(t)h(t, \tau) = \delta(t - \tau), \qquad t_0 \leqslant \tau \leqslant t$$

with

$$h(\tau-, \tau) = \left.\frac{\partial h(t, \tau)}{\partial t}\right|_{t=\tau-} = \cdots = \left.\frac{\partial^{n-1} h(t, \tau)}{\partial t^{n-1}}\right|_{t=\tau-} = 0$$

Notice that

$$h(\tau+, \tau) = \left.\frac{\partial h(t, \tau)}{\partial t}\right|_{t=\tau+} = \cdots = \left.\frac{\partial^{n-3} h(t, \tau)}{\partial t^{n-3}}\right|_{t=\tau+} = \left.\frac{\partial^{n-2} h(t, \tau)}{\partial t^{n-2}}\right|_{t=\tau+} = 0$$

and

$$\left.\frac{\partial^{n-1} h(t, \tau)}{\partial t^{n-1}}\right|_{t=\tau+} = 1$$

Determination of impulse response function Consider the linear time-invariant system defined by

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u, \qquad \mathbf{x}(0-) = \mathbf{0}$$

where $\mathbf{x} = n$ vector (state vector)
$u =$ scalar (forcing function or input function)
$\mathbf{A} = n \times n$ constant matrix
$\mathbf{B} = n \times 1$ constant matrix

Let us obtain the response of the system to the unit impulse input.

For the unit impulse input

$$u(t) = \delta(t)$$

As discussed in Example 6–4 the solution $\mathbf{x}(t)$ is obtained as

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}(0-) + \int_{0-}^{t} e^{\mathbf{A}(t-\tau)}\mathbf{B}\delta(\tau)\, d\tau = e^{\mathbf{A}t}\mathbf{B}, \qquad t \geqslant 0 \qquad (6\text{–}56)$$

Notice that the state space trajectory will exhibit a jump at $t = 0$.

Example 6-8 Obtain the response of the following system:

$$\ddot{x} + (a + b)\dot{x} + abx = u, \qquad x(0-) = \dot{x}(0-) = 0 \qquad (6\text{–}57)$$

where a and b are constants ($a \neq b$) and $u(t) = \delta(t) =$ unit impulse. We shall obtain the response of this system by two methods, (1) the conventional Laplace transform method; (2) the method discussed in this chapter.

(1) The transfer function of the system of Eq. (6–57) is

$$\frac{X(s)}{U(s)} = \frac{1}{s^2 + (a + b)s + ab}$$

The impulse response is, therefore, given by

$$X(s) = \frac{1}{b - a}\left(\frac{1}{s + a} - \frac{1}{s + b}\right)$$

Taking the inverse Laplace transform we obtain

$$x(t) = \frac{1}{a - b}(-e^{-at} + e^{-bt})$$

(2) Equation (6–57) can be rewritten as follows:

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix}=\begin{bmatrix}0 & 1\\ -ab & -(a+b)\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}+\begin{bmatrix}0\\ 1\end{bmatrix}[u] \tag{6–58}$$

where $x_1 = x, \quad x_2 = \dot{x}$

In order to simplify computation of $e^{\mathbf{A}t}$, let us employ the following transformation:

$$\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}1 & 1\\ -a & -b\end{bmatrix}\begin{bmatrix}y_1\\ y_2\end{bmatrix}$$

Equation (6–58) then becomes

$$\begin{bmatrix}\dot{y}_1\\ \dot{y}_2\end{bmatrix}=\begin{bmatrix}-a & 0\\ 0 & -b\end{bmatrix}\begin{bmatrix}y_1\\ y_2\end{bmatrix}+\begin{bmatrix}\frac{-b}{a-b} & \frac{-1}{a-b}\\ \frac{a}{a-b} & \frac{1}{a-b}\end{bmatrix}\begin{bmatrix}0\\ 1\end{bmatrix}[u]$$

$$=\begin{bmatrix}-a & 0\\ 0 & -b\end{bmatrix}\begin{bmatrix}y_1\\ y_2\end{bmatrix}+\begin{bmatrix}\frac{-1}{a-b}\\ \frac{1}{a-b}\end{bmatrix}[u] \tag{6–59}$$

Referring to Eq. (6–56), the solution of Eq. (6–59) to the unit impulse input $\delta(t)$ is given by

$$\mathbf{y}(t) = e^{\mathbf{A}t}\mathbf{B}$$

where

$$e^{\mathbf{A}t}=\begin{bmatrix}e^{-at} & 0\\ 0 & e^{-bt}\end{bmatrix}, \qquad \mathbf{B}=\begin{bmatrix}\frac{-1}{a-b}\\ \frac{1}{a-b}\end{bmatrix}$$

Thus

$$\begin{bmatrix}y_1(t)\\ y_2(t)\end{bmatrix}=\begin{bmatrix}-\frac{e^{-at}}{a-b}\\ \frac{e^{-bt}}{a-b}\end{bmatrix}$$

Hence the solution $\mathbf{x}$ is given by

$$\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}1 & 1\\ -a & -b\end{bmatrix}\begin{bmatrix}-\frac{e^{-at}}{a-b}\\ \frac{e^{-bt}}{a-b}\end{bmatrix}$$

$$=\begin{bmatrix}-\frac{e^{-at}}{a-b}+\frac{e^{-bt}}{a-b}\\ \frac{ae^{-at}}{a-b}-\frac{be^{-bt}}{a-b}\end{bmatrix}$$

and the impulse response of the system is given by

$$x(t) = x_1(t) = \frac{1}{a-b}(-e^{-at} + e^{-bt})$$

We shall next consider determination of the impulse response function of a linear time-varying system by use of an example.

Example 6-9 Consider the following system:

$$\dot{x} + a(t)x = \delta(t-\tau), \qquad x(t<\tau) = 0$$

In this system the unit impulse input is applied at time τ. We shall obtain the impulse response function of this system by two methods, (1) the conventional method, (2) the method discussed in this chapter.

(1) Let us define the impulse response function as $h(t, \tau)$. Then

$$\frac{\partial h(t,\tau)}{\partial t} + a(t)h(t,\tau) = \delta(t-\tau) \tag{6–60}$$

$$0 \leqslant \tau \leqslant t$$

The homogeneous equation

$$\frac{\partial h(t,\tau)}{\partial t} + a(t)h(t,\tau) = 0$$

has a solution

$$\ln[h(t,\tau)] = -\int a(t)dt + \xi(\tau), \qquad t \geqslant \tau$$

where $\xi(\tau)$ is a function of τ. Thus,

$$h(t,\tau) = \eta(\tau)e^{-\int a(t)dt}, \qquad t \geqslant \tau$$

where $\eta(\tau) = e^{\xi(\tau)}$. Since in this problem $x(t<\tau)$ is given as zero, we have

$$h(t,\tau) = 0 \qquad \text{for} \qquad t < \tau$$

Therefore we may write

$$\begin{aligned} h(t,\tau) &= \eta(\tau)e^{-\int a(t)dt}, & t \geqslant \tau\\ &= 0 & t < \tau \end{aligned}$$

Hence

$$\begin{aligned} \frac{\partial h(t,\tau)}{\partial t} &= \delta(t-\tau)\eta(\tau)e^{-\int a(t)dt} - a(t)\eta(\tau)e^{-\int a(t)dt}, & t \geqslant \tau\\ &= 0 & t < \tau \end{aligned} \tag{6–61}$$

Substituting Eq. (6–61) into Eq. (6–60),

$$\delta(t-\tau)\eta(\tau)e^{-\int a(t)dt} = \delta(t-\tau) \tag{6–62}$$

At $t = \tau$ we have

$$\eta(\tau)e^{-\int a(\tau)d\tau} = 1$$

from which

$$\eta(\tau) = e^{\int a(\tau)d\tau}$$

Clearly, Eq. (6–62) is satisfied whenever $\tau \neq t$, since $\delta(t-\tau) = 0$ for $\tau \neq t$. Therefore the impuse response function is obtained as follows:

$$\begin{aligned} h(t,\tau) &= e^{\int a(\tau)d\tau - \int a(t)dt}\\ &= e^{\int_{t_0}^{\tau} a(s)ds - \int_{t_0}^{t} a(s)ds} = e^{-\int_{\tau}^{t} a(s)ds}, & t \geqslant \tau\\ &= 0 & t < \tau \end{aligned}$$

(2) The system equation can be written as

$$\dot{x} = -a(t)x + \delta(t-\tau), \qquad x(t<\tau) = 0$$

The impulse response function can be expressed as

$$h(t, \tau) = \phi(t, t_0)\psi(\tau, t_0)$$

where $\phi(t, t_0)$ is the state transition function (one-dimensional state transition matrix) and $\psi(\tau, t_0) = \phi^{-1}(\tau, t_0)$. Since $a(t)$ is a scalar, $a(t)$ and $\int_{t_0}^{t} a(s)ds$ commute for all t. Hence

$$\phi(t, t_0) = e^{-\int_{t_0}^{t} a(s)ds}$$

Then

$$\psi(\tau, t_0) = \phi^{-1}(\tau, t_0) = e^{\int_{t_0}^{\tau} a(s)ds}$$

The impulse response function is then obtained as

$$h(t, \tau) = e^{-\int_{t_0}^{t} a(s)ds + \int_{t_0}^{\tau} a(s)ds} = e^{-\int_{\tau}^{t} a(s)ds}, \qquad t \geqslant \tau$$

We shall next consider the response of the system studied in Example 6–9 for an arbitrary input.

Example 6-10 Obtain the response of the following system:

$$\dot{x} + a(t)x = u, \qquad x(t < t_0) = 0$$

where u is an arbitrary input.

The response is given by

$$\begin{aligned} x(t) &= \int_{t_0}^{t} h(t, \tau)u(\tau)d\tau \\ &= \int_{t_0}^{t} e^{-\int_{t_0}^{t} a(s)ds + \int_{t_0}^{\tau} a(s)ds} u(\tau)d\tau \\ &= e^{-\int_{t_0}^{t} a(s)ds} \int_{t_0}^{t} e^{\int_{t_0}^{\tau} a(s)ds} u(\tau)d\tau, \qquad t \geqslant t_0 \end{aligned}$$

6-7 VECTOR MATRIX DIFFERENCE EQUATIONS

Mathematical representations of discrete-time systems are very often given in the form of vector matrix difference equations. Formal solutions of vector matrix difference equations can be obtained by techniques similar to those discussed in Sections 6–2 and 6–3 for linear vector matrix differential equations.

In analyzing discrete-time systems, without loss of generality, we can assume the sampling period T to be equal to unity. In order to simplify notation we shall employ this assumption in the following analysis unless otherwise stated.

The equation with which we shall deal in this section is of the following form:

$$\mathbf{x}(k + 1) = \mathbf{G}(k)\mathbf{x}(k) + \mathbf{H}(k)\mathbf{u}(k) \qquad (6\text{–}63)$$

where $\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$k =$ integer which indicates kth sampling instant

$\mathbf{G}(k) = n \times n$ matrix
$\mathbf{H}(k) = n \times r$ matrix

Equation (6–63) is equivalent to a set of n difference equations of the following form[1]:

$$x_i(k + 1) = \sum_{j=1}^{n} g_{ij}(k)x_j(k) + \sum_{j=1}^{r} h_{ij}(k)u_j(k) \qquad (i = 1, 2, \ldots, n)$$

We shall first present solutions of homogeneous vector matrix difference equations and then give solutions of nonhomogeneous vector matrix difference equations. Finally, we shall discuss discretization of vector matrix differential equations.

Homogeneous vector matrix difference equation Consider the following linear homogeneous vector matrix difference equation:

$$\mathbf{x}(k + 1) = \mathbf{G}(k)\mathbf{x}(k), \qquad \mathbf{x}(h) = \mathbf{x}_0 \qquad (6\text{–}64)$$
$$(k = h, h + 1, h + 2, \ldots; h = \text{integer})$$

Similar to the case of linear vector matrix differential equations, an $n \times n$ matrix $\mathbf{X}$ such that its n columns consist of n linearly independent solutions of Eq. (6–64) is called a *fundamental matrix*. It is nonsingular and satisfies the matrix difference equation

$$\mathbf{X}(k + 1) = \mathbf{G}(k)\,\mathbf{X}(k)$$

Notice that n columns of a fundamental matrix constitute a basis of the n-dimensional vector space.

Consider the unique fundamental matrix $\Phi(k, h)$ which satisfies

$$\Phi(k + 1, h) = \mathbf{G}(k)\Phi(k, h), \qquad \Phi(h, h) = \mathbf{I}$$

Any fundamental matrix $\mathbf{X}(k)$ can be written as $\Phi(k, h)\mathbf{C}$ where $\mathbf{C}$ is an $n \times n$ nonsingular matrix. Notice that $\Phi(k, h)$ has the property that

$$\begin{aligned} \Phi(h, h) &= \mathbf{I} \\ \Phi(h + 1, h) &= \mathbf{G}(h) \\ \Phi(h + 2, h) &= \mathbf{G}(h + 1)\,\mathbf{G}(h) \\ \Phi(h + 3, h) &= \mathbf{G}(h + 2)\,\mathbf{G}(h + 1)\,\mathbf{G}(h) \\ &\cdots \qquad \cdots \qquad \cdots \\ \Phi(k, h) &= \mathbf{G}(k - 1)\mathbf{G}(k - 2) \cdots \mathbf{G}(h + 1)\,\mathbf{G}(h) \end{aligned}$$

Using $\Phi(k, h)$ the solution of Eq. (6–64) is given by

$$\mathbf{x}(k) = \Phi(k, h)\mathbf{x}(h) = \Phi(k, h)\,\mathbf{x}_0$$

---

[1]The solution of this equation is well known; see, for example (*6–8*). [Numbers in parentheses refer to list of references at the end of the chapter.]

If $\mathbf{G}(k)$ is an $n \times n$ constant matrix, which we shall denote by $\mathbf{G}$, then by choosing $h = 0$ and writing $\Phi(k, h)$ as $\Phi(k)$ we obtain

$$\Phi(k) = \mathbf{G}^k$$

$\Phi(k)$ is the unique fundamental matrix satisfying

$$\Phi(k+1) = \mathbf{G}\Phi(k), \qquad \Phi(0) = \mathbf{I}$$

The solution to

$$\mathbf{x}(k+1) = \mathbf{G}\mathbf{x}(k), \qquad \mathbf{x}(0) = \mathbf{x}_0$$

is therefore given by

$$\mathbf{x}(k) = \mathbf{G}^k\mathbf{x}_0$$

Note that if $\mathbf{G}$ is a diagonal matrix, or

$$\mathbf{G} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

then the unique fundamental matrix $\Phi(k)$ is given by

$$\Phi(k) = \mathbf{G}^k = \text{diag}\,(\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k)$$

Nonhomogeneous vector matrix difference equation Consider the nonhomogeneous vector matrix difference equation

$$\mathbf{x}(k+1) = \mathbf{G}(k)\mathbf{x}(k) + \mathbf{H}(k)\mathbf{u}(k), \qquad \mathbf{x}(h) = \mathbf{x}_0 \tag{6-65}$$
$$(k = h, h+1, h+2, \ldots; h = \text{integer})$$

The solution of Eq. (6–65) is given by

$$\mathbf{x}(k) = \Phi(k, h)\mathbf{x}_0 + \Phi(k, h)\sum_{j=h+1}^{k}\Phi^{-1}(j, h)\mathbf{H}(j-1)\mathbf{u}(j-1) \qquad (k > h) \tag{6-66}$$

where $\Phi(k, h)$ is the unique fundamental matrix satisfying

$$\Phi(k+1, h) = \mathbf{G}(k)\Phi(k, h), \qquad \Phi(h, h) = \mathbf{I} \qquad (k = h, h+1, h+2, \ldots)$$

and $\mathbf{H}(k)$ and $\mathbf{u}(k)$ are zero for $k < h$.

We can prove this by direct substitution. For $k > h$, we obtain from Eq. (6–66)

$$\mathbf{x}(k+1) = \Phi(k+1, h)\mathbf{x}_0 + \Phi(k+1, h)\sum_{j=h+1}^{k+1}\Phi^{-1}(j, h)\mathbf{H}(j-1)\mathbf{u}(j-1) \tag{6-67}$$

Noting that

$$\Phi(k, h) = \mathbf{G}(k-1)\mathbf{G}(k-2)\cdots\mathbf{G}(h)$$

Eq. (6–67) can be expressed as follows:

$$\begin{aligned}\mathbf{x}(k+1) &= \mathbf{G}(k)\Phi(k, h)\mathbf{x}_0 + \mathbf{G}(k)\Phi(k, h)\sum_{j=h+1}^{k}\Phi^{-1}(j, h)\mathbf{H}(j-1)\mathbf{u}(j-1) \\ &\quad + \Phi(k+1, h)\Phi^{-1}(k+1, h)\mathbf{H}(k)\mathbf{u}(k) \\ &= \mathbf{G}(k)\mathbf{x}(k) + \mathbf{H}(k)\mathbf{u}(k)\end{aligned}$$

Thus, we have shown that Eq. (6–66) is the solution of Eq. (6–65). From Eq. (6–66) we notice that the solution to the nonhomogeneous vector matrix difference equation consists of two parts: the first is the homogeneous solution, the second is a particular solution.

For the linear time-invariant vector matrix difference equation

$$\mathbf{x}(k+1) = \mathbf{G}\mathbf{x}(k) + \mathbf{H}\mathbf{u}(k), \qquad \mathbf{x}(0) = \mathbf{x}_0$$

the solution, Eq. (6–66), is simplified as follows: For $k = 1, 2, 3, \ldots$

$$\begin{aligned}\mathbf{x}(k) &= \mathbf{G}^k\mathbf{x}_0 + \mathbf{G}^k\sum_{j=1}^{k}\mathbf{G}^{-j}\mathbf{H}\mathbf{u}(j-1) \\ &= \mathbf{G}^k\mathbf{x}_0 + \sum_{j=1}^{k}\mathbf{G}^{k-j}\mathbf{H}\mathbf{u}(j-1)\end{aligned}$$

This solution corresponds to Eq. (6–21).

Discretization of vector matrix differential equation A continuous-time system which is described by a vector matrix differential equation can be approximated by a vector matrix difference equation. Such approximation is often employed in practice. In particular, if we wish to compute system responses to various inputs by the use of a digital computer, we must convert a continuous-time system to a discrete-time system. The discrete-time solutions must be valid at equally spaced sampling instants. In discretizing the continuous-time system we assume the inputs or the forcing functions to change only at these same instants.

The following material presents procedures for deriving the difference equation model from the differential equation model. Consider a system whose dynamic behavior is governed by the following vector matrix differential equation:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \tag{6-68}$$

where

$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{A} = n \times n$ constant matrix
$\mathbf{B} = n \times r$ constant matrix

If the components of $\mathbf{x}$ are measured only every T seconds, the system is a discrete-time system. In the following, in order to clarify the analysis we shall use the notation kT and $(k+1)T$ instead of k and $(k+1)$.

The linear discrete-time system can be written in the following form of the vector matrix difference equation relating the state $\mathbf{x}$ at the sampling instant $(k+1)T$ to its state at kT, T seconds before:

$$\begin{aligned}\mathbf{x}((k+1)T) &= \mathbf{G}(T)[\mathbf{x}(kT) + \mathbf{F}(T)\mathbf{u}(kT)] \\ &= \mathbf{G}(T)\mathbf{x}(kT) + \mathbf{H}(T)\mathbf{u}(kT)\end{aligned} \tag{6-69}$$

where $\mathbf{G}(T)$ is an $n \times n$ constant matrix whose elements are functions of the sampling period T, $\mathbf{F}(T)$ is an $n \times r$ constant matrix whose elements are also functions of T, and $\mathbf{H}(T) = \mathbf{G}(T)\mathbf{F}(T)$. The recursive relationship given by Eq. (6–69) allows us to compute the state at time $(k+1)T$ from the state at time kT and the control vector at time kT. To determine $\mathbf{G}(T)$ and $\mathbf{F}(T)$ we may use the solution of Eq. (6–68), or

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}(0) + e^{\mathbf{A}t}\int_0^t e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}(\tau)\,d\tau$$

We assume that all the components of $\mathbf{u}(t)$ are constant over the interval between any two consecutive sampling instants, or $\mathbf{u}(t) = \mathbf{u}(kT)$ for the kth sampling period. Then, $\mathbf{x}((k+1)T)$ can be obtained as follows:

$$\begin{aligned}\mathbf{x}((k+1)T) &= e^{\mathbf{A}T}\mathbf{x}(kT) + e^{\mathbf{A}T}\int_0^T e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}(kT)\,d\tau \\ &= e^{\mathbf{A}T}\mathbf{x}(kT) + \int_0^T e^{\mathbf{A}\lambda}\mathbf{B}\mathbf{u}(kT)\,d\lambda\end{aligned}$$

where $\lambda = T - \tau$. (For the derivation of this last equation, refer to Prob. A-6-28.) If we define

$$\mathbf{G}(T) = e^{\mathbf{A}T}$$

$$\mathbf{F}(T) = \left(\int_0^T e^{-\mathbf{A}\tau}\,d\tau\right)\mathbf{B} = \left(e^{-\mathbf{A}T}\int_0^T e^{\mathbf{A}\tau}\,d\tau\right)\mathbf{B}$$

and

$$\mathbf{H}(T) = \mathbf{G}(T)\mathbf{F}(T) = \left(\int_0^T e^{\mathbf{A}\tau}\,d\tau\right)\mathbf{B}$$

then

$$\begin{aligned}\mathbf{x}((k+1)T) &= e^{\mathbf{A}T}\mathbf{x}(kT) + e^{\mathbf{A}T}\int_0^T e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}(kT)\,d\tau \\ &= e^{\mathbf{A}T}\left[\mathbf{x}(kT) + \int_0^T e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}(kT)\,d\tau\right] \\ &= \mathbf{G}(T)[\mathbf{x}(kT) + \mathbf{F}(T)\mathbf{u}(kT)] \\ &= \mathbf{G}(T)\mathbf{x}(kT) + \mathbf{H}(T)\mathbf{u}(kT)\end{aligned}$$

which is Eq. (6–69). Thus Eq. (6–68) reduces to Eq. (6–69). [Notice that Eq. (6–69) represents the solution of Eq. (6–68) only at discrete points of time $t = kT$; $k = 0, 1, 2, \ldots$.]

Example 6-11 Given the following vector matrix differential equation:

$$\begin{bmatrix}\dot{x}_1 \\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}0 & 1 \\ 0 & -2\end{bmatrix}\begin{bmatrix}x_1 \\ x_2\end{bmatrix} + \begin{bmatrix}0 \\ 1\end{bmatrix}[u]$$

obtain an equivalent discrete-time representation of the system.

An equivalent discrete-time system is given by

$$\mathbf{x}((k+1)T) = \mathbf{G}(T)\mathbf{x}(kT) + \mathbf{H}(T)u(kT)$$

where $\mathbf{G}(T)$ is obtained as

$$\mathbf{G}(T) = e^{\mathbf{A}T} = \begin{bmatrix}1 & \frac{1}{2}(1 - e^{-2T}) \\ 0 & e^{-2T}\end{bmatrix}$$

and $\mathbf{H}(T)$ is obtained as

$$\begin{aligned}\mathbf{H}(T) &= \left(\int_0^T e^{\mathbf{A}\tau}\,d\tau\right)\mathbf{B} \\ &= \int_0^T \begin{bmatrix}1 & \frac{1}{2}(1 - e^{-2\tau}) \\ 0 & e^{-2\tau}\end{bmatrix} d\tau \begin{bmatrix}0 \\ 1\end{bmatrix} \\ &= \begin{bmatrix}\frac{1}{2}\left(T + \frac{e^{-2T} - 1}{2}\right) \\ \frac{1}{2}(1 - e^{-2T})\end{bmatrix}\end{aligned}$$

Hence

$$\begin{bmatrix}x_1((k+1)T) \\ x_2((k+1)T)\end{bmatrix} = \begin{bmatrix}1 & \frac{1}{2}(1 - e^{-2T}) \\ 0 & e^{-2T}\end{bmatrix}\begin{bmatrix}x_1(kT) \\ x_2(kT)\end{bmatrix} + \frac{1}{2}\begin{bmatrix}T + \frac{e^{-2T} - 1}{2} \\ 1 - e^{-2T}\end{bmatrix}[u(kT)]$$

In computing $\mathbf{G}$ and $\mathbf{H}$ we may use the Laplace transform method. Assuming $\mathbf{u}(t)$ to be constant over each sampling period and shifting the origin of the time axis by putting $\tau = t - kT$, we obtain the Laplace transform of Eq. (6–68) for the period $0 \leqslant \tau \leqslant T$ as follows:

$$s\mathbf{X}(s) - \mathbf{x}(0) = \mathbf{A}\mathbf{X}(s) + \frac{1}{s}\mathbf{B}\mathbf{u}(0)$$

where

$$\mathbf{X}(s) = \mathscr{L}[\mathbf{x}(\tau)]$$

The inverse Laplace transform of $\mathbf{X}(s)$ then yields the solution of Eq. (6–68). Namely,

$$\begin{aligned}\mathbf{x}(\tau) &= \mathscr{L}^{-1}[\mathbf{X}(s)] \\ &= \mathscr{L}^{-1}[(s\mathbf{I} - \mathbf{A})^{-1}]\mathbf{x}(0) + \mathscr{L}^{-1}\left[\frac{1}{s}(s\mathbf{I} - \mathbf{A})^{-1}\right]\mathbf{B}\mathbf{u}(0)\end{aligned}$$

where

$$(s\mathbf{I} - \mathbf{A})^{-1} = \frac{1}{|s\mathbf{I} - \mathbf{A}|}\,[\text{cofactor of } |s\mathbf{I} - \mathbf{A}|]$$

$\mathbf{G}(\tau)$ and $\mathbf{H}(\tau)$ are thus obtained as

$$\mathbf{G}(\tau) = \mathscr{L}^{-1}[(s\mathbf{I} - \mathbf{A})^{-1}]$$

and

$$\mathbf{H}(\tau) = \mathscr{L}^{-1}\left[\frac{1}{s}(s\mathbf{I} - \mathbf{A})^{-1}\right]\mathbf{B}$$

or more conveniently

$$\mathbf{H}(\tau) = \int_0^\tau \mathscr{L}^{-1}[(s\mathbf{I} - \mathbf{A})^{-1}]\mathbf{B}\,d\sigma$$
$$= \int_0^\tau \mathbf{G}(\sigma)\mathbf{B}\,d\sigma$$

To obtain $\mathbf{G}(T)$ and $\mathbf{H}(T)$, τ is set equal to T.

Example 6-12 Consider the following system:

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix} = \begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ -6 & -11 & -6\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix} + \begin{bmatrix}0\\ 0\\ 1\end{bmatrix}[u]$$

Let us obtain the explicit forms of $\mathbf{G}(\tau)$ and $\mathbf{H}(\tau)$, where

$$\mathbf{x}(kT + \tau) = \mathbf{G}(\tau)\mathbf{x}(kT) + \mathbf{H}(\tau)u(kT)$$

The elements $g_{ij}(\tau)$ correspond to the elements of

$$\mathscr{L}^{-1}[(s\mathbf{I} - \mathbf{A})^{-1}]$$

where

$$\mathbf{A} = \begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ -6 & -11 & -6\end{bmatrix}$$

Thus

$$s\mathbf{I} - \mathbf{A} = \begin{bmatrix}s & -1 & 0\\ 0 & s & -1\\ 6 & 11 & s+6\end{bmatrix}$$

and

$$(s\mathbf{I} - \mathbf{A})^{-1} = \frac{1}{\Delta}\begin{bmatrix}s(s+6)+11 & s+6 & 1\\ -6 & s(s+6) & s\\ -6s & -11s-6 & s^2\end{bmatrix}$$

where $\Delta = (s+1)(s+2)(s+3)$

Taking the inverse Laplace transform of each element of $(s\mathbf{I} - \mathbf{A})^{-1}$ we obtain

$$\mathbf{G}(\tau) = \begin{bmatrix}3e^{-\tau} - 3e^{-2\tau} + e^{-3\tau} & \frac{5}{2}e^{-\tau} - 4e^{-2\tau} + \frac{3}{2}e^{-3\tau} & \frac{1}{2}e^{-\tau} - e^{-2\tau} + \frac{1}{2}e^{-3\tau}\\ -3e^{-\tau} + 6e^{-2\tau} - 3e^{-3\tau} & -\frac{5}{2}e^{-\tau} + 8e^{-2\tau} - \frac{9}{2}e^{-3\tau} & -\frac{1}{2}e^{-\tau} + 2e^{-2\tau} - \frac{3}{2}e^{-3\tau}\\ 3e^{-\tau} - 12e^{-2\tau} + 9e^{-3\tau} & \frac{5}{2}e^{-\tau} - 16e^{-2\tau} + \frac{27}{2}e^{-3\tau} & \frac{1}{2}e^{-\tau} - 4e^{-2\tau} + \frac{9}{2}e^{-3\tau}\end{bmatrix}$$

The elements of $\mathbf{H}(\tau)$ are found from the expression

$$\mathbf{H}(\tau) = \int_0^\tau \mathbf{G}(\sigma)\mathbf{B}\,d\sigma$$

where

$$\mathbf{B} = \begin{bmatrix}0\\ 0\\ 1\end{bmatrix}$$

Thus,

$$\mathbf{H}(\tau) = \begin{bmatrix}-\frac{1}{2}e^{-\tau} + \frac{1}{2}e^{-2\tau} - \frac{1}{6}e^{-3\tau} + \frac{1}{6}\\ \frac{1}{2}e^{-\tau} - e^{-2\tau} + \frac{1}{2}e^{-3\tau}\\ -\frac{1}{2}e^{-\tau} + 2e^{-2\tau} - \frac{3}{2}e^{-3\tau}\end{bmatrix}$$

EXERCISE PROBLEMS AND SOLUTIONS

Problem A-6-1 Prove that if $x_1(t), x_2(t), \ldots, x_m(t)$ are m solutions of

$$L(x) = \overset{(n)}{x} + a_1(t)\overset{(n-1)}{x} + \cdots + a_{n-1}(t)\dot{x} + a_n(t)x = 0$$

then $c_1x_1(t) + c_2x_2(t) + \cdots + c_mx_m(t)$, where $c_1, c_2, \ldots, c_m$ are arbitrary constants, is also a solution of $L(x) = 0$.

Solution The linear operation $L(x)$ has the property that for arbitrary constants α and β and arbitrary functions f and g

$$L(\alpha f + \beta g) = \alpha L(f) + \beta L(g)$$

Hence

$$L(c_1x_1 + c_2x_2 + \cdots + c_mx_m) = c_1L(x_1) + c_2L(x_2) + \cdots + c_mL(x_m)$$

By assumption, $x_1, x_2, \ldots, x_m$ are solutions of $L(x) = 0$. Therefore, $L(x_1) = L(x_2) = \cdots = L(x_m) = 0$. Hence

$$L(c_1x_1 + c_2x_2 + \cdots + c_mx_m) = 0$$

This implies that $c_1x_1 + c_2x_2 + \cdots + c_mx_m$ is a solution of $L(x) = 0$.

Problem A-6-2 Show that a general solution of a linear differential equation $L(x) = X(t)$ is given as a sum of n linearly independent solutions of $L(x) = 0$ and a particular solution of $L(x) = X(t)$.

Solution Let $x_0(t)$ be a particular solution of $L(x) = X(t)$ and $y(t)$ be a general solution of $L(y) = 0$. Then

$$L(y + x_0) = L(y) + L(x_0) = X(t)$$

Thus, $y + x_0$ is a solution of $L(x) = X(t)$. Since $y(t)$ involves n arbitrary constants, $y + x_0$ also involves n arbitrary constants. Since no singular solution exists in this linear system, $y + x_0$ is a general solution of $L(x) = X(t)$.

Problem A-6-3 Consider a system defined by

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t), \qquad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{6-70}$$

where the eigenvalues of $\mathbf{A}$ are distinct. Show that a fundamental matrix can be given by

$$\mathbf{X}(t) = \begin{bmatrix}e^{\lambda_1(t-t_0)}c_{11} & e^{\lambda_2(t-t_0)}c_{12} & \cdots & e^{\lambda_n(t-t_0)}c_{1n}\\ e^{\lambda_1(t-t_0)}c_{21} & e^{\lambda_2(t-t_0)}c_{22} & \cdots & e^{\lambda_n(t-t_0)}c_{2n}\\ \vdots & \vdots & & \vdots\\ e^{\lambda_1(t-t_0)}c_{n1} & e^{\lambda_2(t-t_0)}c_{n2} & \cdots & e^{\lambda_n(t-t_0)}c_{nn}\end{bmatrix}$$

and the unique fundamental matrix $\Phi(t)$ which satisfies

$$\dot{\Phi}(t) = \mathbf{A}\Phi(t), \qquad \Phi(t_0) = \mathbf{I}$$

can be given by

$$\Phi(t) = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix} \begin{bmatrix} e^{\lambda_1(t-t_0)} & & & 0 \\ & e^{\lambda_2(t-t_0)} & & \\ & & \cdot & \\ & & & \cdot \\ 0 & & & e^{\lambda_n(t-t_0)} \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix}^{-1}$$

Solution Let us assume the solution of Eq. (6–70) to be

$$\mathbf{x}(t) = e^{\lambda(t-t_0)}\mathbf{c} \tag{6–71}$$

Then,

$$\dot{\mathbf{x}}(t) = \lambda e^{\lambda(t-t_0)}\mathbf{c} \tag{6–72}$$

From Eq. (6–71) and Eq. (6–72), we obtain

$$\lambda e^{\lambda(t-t_0)}\mathbf{c} = \mathbf{A}e^{\lambda(t-t_0)}\mathbf{c}$$

or

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{c} = \mathbf{0}$$

which may be written as

$$\begin{bmatrix} a_{11} - \lambda & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & a_{23} & \cdots & a_{2n} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} - \lambda \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \cdot \\ \cdot \\ \cdot \\ c_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} \tag{6–73}$$

Equation (6–73) has a nontrivial solution $\mathbf{c} \neq \mathbf{0}$ if and only if

$$|\mathbf{A} - \lambda\mathbf{I}| = 0$$

This is an nth-degree polynomial in λ and is the characteristic equation of the given system. Notice that for each λ_i there corresponds an eigenvector $\mathbf{c}_i$. The coordinates of the vectors $\mathbf{c}_i$ can be obtained by solving the following n simultaneous equations with $\lambda = \lambda_i$:

$$\begin{aligned} (a_{11} - \lambda_i)c_{1i} + a_{12}c_{2i} + a_{13}c_{3i} + \cdots + a_{1n}c_{ni} &= 0 \\ a_{21}c_{1i} + (a_{22} - \lambda_i)c_{2i} + a_{23}c_{3i} + \cdots + a_{2n}c_{ni} &= 0 \\ \cdots \quad \cdots \quad \cdots \\ a_{n1}c_{1i} + a_{n2}c_{2i} + a_{n3}c_{3i} + \cdots + (a_{nn} - \lambda_i)c_{ni} &= 0 \end{aligned}$$
$$(i = 1, 2, 3, \ldots, n)$$

Then, the vectors

$$\mathbf{x}_i(t) = e^{\lambda_i(t-t_0)}\mathbf{c}_i = \begin{bmatrix} e^{\lambda_i(t-t_0)}c_{1i} \\ e^{\lambda_i(t-t_0)}c_{2i} \\ \cdot \\ \cdot \\ \cdot \\ e^{\lambda_i(t-t_0)}c_{ni} \end{bmatrix}$$

are n solutions of Eq. (6–70). Since the $\mathbf{x}_i(t)$ are linearly independent over any interval, the fundamental matrix is given by

$$\mathbf{X}(t) = \begin{bmatrix} e^{\lambda_1(t-t_0)}c_{11} & e^{\lambda_2(t-t_0)}c_{12} & \cdots & e^{\lambda_n(t-t_0)}c_{1n} \\ e^{\lambda_1(t-t_0)}c_{21} & e^{\lambda_2(t-t_0)}c_{22} & \cdots & e^{\lambda_n(t-t_0)}c_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ e^{\lambda_1(t-t_0)}c_{n1} & e^{\lambda_2(t-t_0)}c_{n2} & \cdots & e^{\lambda_n(t-t_0)}c_{nn} \end{bmatrix}$$

$$= \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix} \begin{bmatrix} e^{\lambda_1(t-t_0)} & & & 0 \\ & e^{\lambda_2(t-t_0)} & & \\ & & \cdot & \\ & & & \cdot \\ 0 & & & e^{\lambda_n(t-t_0)} \end{bmatrix}$$

Since the unique fundamental matrix $\Phi(t)$ which satisfies

$$\dot{\Phi}(t) = \mathbf{A}\Phi(t), \qquad \Phi(t_0) = \mathbf{I}$$

where $\mathbf{A}$ is an $n \times n$ constant matrix, can be written as $\mathbf{X}(t)$ times a constant nonsingular matrix, we obtain the following relationship:

$$\Phi(t) = \mathbf{X}(t)\mathbf{X}^{-1}(t_0)$$

$$= \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix} \begin{bmatrix} e^{\lambda_1(t-t_0)} & & & 0 \\ & e^{\lambda_2(t-t_0)} & & \\ & & \cdot & \\ & & & \cdot \\ 0 & & & e^{\lambda_n(t-t_0)} \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix}^{-1}$$

Problem A-6-4 Consider a system described by the following transfer function:

$$\frac{X(s)}{U(s)} = \frac{K}{(s+a)(s+b)(s+c)}$$

Find a few sets of basis functions.

Solution Examples of a few sets of basis functions are

(1) $\dfrac{1}{s+a}, \quad \dfrac{1}{s+b}, \quad \dfrac{1}{s+c}$

(2) $\dfrac{1}{(s+a)(s+b)(s+c)}, \quad \dfrac{1}{(s+a)(s+b)}, \quad \dfrac{1}{(s+a)}$

(3) $\dfrac{s^2}{(s+a)(s+b)(s+c)}, \quad \dfrac{s}{(s+a)(s+b)(s+c)}, \quad \dfrac{1}{(s+a)(s+b)(s+c)}$

Problem A-6-5 Are the following two systems equivalent?

(1) $\dfrac{X(s)}{U(s)} = \dfrac{(s+b)}{(s+a)(s+b)}$

(2) $\dfrac{X(s)}{U(s)} = \dfrac{(s+c)}{(s+a)(s+c)} \qquad (b \neq c)$

Solution These two systems have the same transfer function, if we cancel the common terms in the numerator and denominator, but they have different basis functions. The basis functions of system (1) are $1/(s+a)$ and $1/(s+b)$ (or their linear combinations); those of system (2) are $1/(s+a)$ and $1/(s+c)$ (or their linear combinations.) Two systems are equivalent if the transfer functions are the same and if the basis functions of one system are linear combinations of the basis functions of the other system, and vice-versa. The present systems are, therefore, not equivalent.

Problem A-6-6 The elements of the fundamental matrix $\Phi(t)$ with $\Phi(0) = \mathbf{I}$ may be generated by the use of summers and integrators. Consider the following second-order system:

$$\ddot{x} + 2\zeta\omega_n\dot{x} + \omega_n^2 x = u$$

which can be rewritten as

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}0 & 1\\ -\omega_n^2 & -2\zeta\omega_n\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + \begin{bmatrix}0\\ 1\end{bmatrix}[u] \tag{6-74}$$

where $x_1 = x$ and $x_2 = \dot{x}$

Draw analog computer diagrams for the given system and indicate where the elements of the fundamental matrix $\Phi(t)$ may appear.

Solution The elements of $\Phi(t)$ can be determined from

$$\begin{bmatrix}\dot{\phi}_{11} & \dot{\phi}_{12}\\ \dot{\phi}_{21} & \dot{\phi}_{22}\end{bmatrix} = \begin{bmatrix}0 & 1\\ -\omega_n^2 & -2\zeta\omega_n\end{bmatrix}\begin{bmatrix}\phi_{11} & \phi_{12}\\ \phi_{21} & \phi_{22}\end{bmatrix}$$

where

$$\begin{bmatrix}\phi_{11}(0) & \phi_{12}(0)\\ \phi_{21}(0) & \phi_{22}(0)\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$

Hence we obtain

$$\phi_{11} = \int_0^t \phi_{21}\,dt + 1$$

$$\phi_{12} = \int_0^t \phi_{22}\,dt$$

$$\phi_{21} = \int_0^t (-\omega_n^2\phi_{11} - 2\zeta\omega_n\phi_{21})\,dt$$

$$\phi_{22} = \int_0^t (-\omega_n^2\phi_{12} - 2\zeta\omega_n\phi_{22})\,dt + 1$$

Figures 6-4(A) and (B) are analog computer diagrams showing where the elements of the fundamental matrix $\Phi(t)$ will appear.

Problem A-6-7 Consider the system described by

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}0 & 1\\ -1 & t\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}$$

Obtain the fundamental matrix $\Phi(t, 0)$ with $\Phi(0, 0) = \mathbf{I}$.

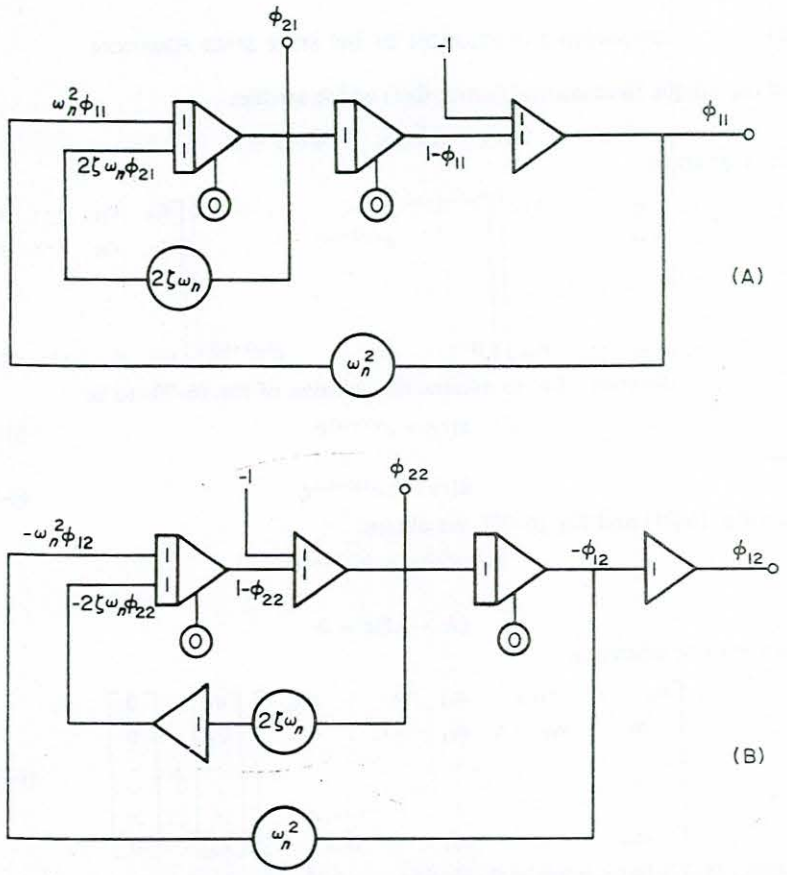


Figure 6-4 Analog computer diagrams for second order system defined by Eq. (6-74). (A) Diagram indicating ϕ_{11} and ϕ_{21}; (B) diagram indicating ϕ_{12} and ϕ_{22}.

Solution From the system equation $\dot{x}_2$ is given by

$$\dot{x}_2 = -x_1 + tx_2 \tag{6-75}$$

Differentiating Eq. (6-75) with respect to t,

$$\ddot{x}_2 = -\dot{x}_1 + t\dot{x}_2 + x_2 = t\dot{x}_2$$

Let us put $v = \dot{x}_2$. Then

$$\dot{v} - tv = 0$$

Solving this last equation

$$v = c_1 e^{\int_0^t \tau\,d\tau} = c_1 e^{(1/2)t^2}$$

Thus

$$x_2(t) = \int v\,dt = c_1\int_0^t e^{(1/2)\tau^2}\,d\tau + x_2(0)$$

$$x_1(t) = \int x_2\,dt = c_1 \int_0^t \int_0^\xi e^{(1/2)\tau^2}\,d\tau\,d\xi + tx_2(0) + x_1(0)$$

Since at $t = 0$

$$\dot{x}_1(0) + x_1(0) = 0$$

it follows that

$$c_1 = -x_1(0)$$

Thus we obtain

$$x_1(t) = x_1(0)\left[1 - \int_0^t \int_0^\xi e^{(1/2)\tau^2}\,d\tau\,d\xi\right] + tx_2(0)$$

$$x_2(t) = -x_1(0) \int_0^t e^{(1/2)\tau^2}\,d\tau + x_2(0)$$

Hence,

$$\Phi(t, 0) = \begin{bmatrix} 1 - \int_0^t \int_0^\xi e^{(1/2)\tau^2}\,d\tau\,d\xi & t \\ -\int_0^t e^{(1/2)\tau^2}\,d\tau & 1 \end{bmatrix}$$

Problem A-6-8 Obtain the fundamental matrix $\Phi(t)$ with $\Phi(0) = \mathbf{I}$ of the system

$$\ddot{x} = ax$$

where a is a constant.

Solution Let

$$x_1 = x$$
$$x_2 = \dot{x}$$
$$x_3 = \ddot{x}$$

then the system equation becomes

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ a & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{A} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Computing $\mathbf{A}^2, \mathbf{A}^3, \ldots$, we obtain

$$\mathbf{A}^2 = \begin{bmatrix} 0 & 0 & 1 \\ a & 0 & 0 \\ 0 & a & 0 \end{bmatrix}, \qquad \mathbf{A}^3 = \begin{bmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & a \end{bmatrix}$$

$$\mathbf{A}^4 = \begin{bmatrix} 0 & a & 0 \\ 0 & 0 & a \\ a^2 & 0 & 0 \end{bmatrix}, \qquad \mathbf{A}^5 = \begin{bmatrix} 0 & 0 & a \\ a^2 & 0 & 0 \\ 0 & a^2 & 0 \end{bmatrix}, \qquad \mathbf{A}^6 = \begin{bmatrix} a^2 & 0 & 0 \\ 0 & a^2 & 0 \\ 0 & 0 & a^2 \end{bmatrix}$$

If we define $\Phi(t) = e^{\mathbf{A}t} = [\mathbf{x}_1 | \mathbf{x}_2 | \mathbf{x}_3]$, then the column vectors $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ are obtained as

$$\mathbf{x}_1 = \begin{bmatrix} 1 + \frac{a}{3!}t^3 + \frac{a^2}{6!}t^6 + \cdots \\ \frac{a}{2!}t^2 + \frac{a^2}{5!}t^5 + \frac{a^3}{8!}t^8 + \cdots \\ at + \frac{a^2}{4!}t^4 + \frac{a^3}{7!}t^7 + \cdots \end{bmatrix}$$

$$\mathbf{x}_2 = \begin{bmatrix} t + \frac{a}{4!}t^4 + \frac{a^2}{7!}t^7 + \cdots \\ 1 + \frac{a}{3!}t^3 + \frac{a^2}{6!}t^6 + \cdots \\ \frac{a}{2!}t^2 + \frac{a^2}{5!}t^5 + \frac{a^3}{8!}t^8 + \cdots \end{bmatrix}$$

$$\mathbf{x}_3 = \begin{bmatrix} \frac{1}{2!}t^2 + \frac{a}{5!}t^5 + \frac{a^2}{8!}t^8 + \cdots \\ t + \frac{a}{4!}t^4 + \frac{a^2}{7!}t^7 + \cdots \\ 1 + \frac{a}{3!}t^3 + \frac{a^2}{6!}t^6 + \cdots \end{bmatrix}$$

Problem A-6-9 (6–4) Consider a linear system described by

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$$

where $\mathbf{x}$ is an n vector and $\mathbf{A}$ is an $n \times n$ constant matrix. Utilizing the method of successive approximations, an approximate solution of the system equation can be given by

$$\mathbf{x}_n(t) = \int_0^t \mathbf{A}\mathbf{x}_{n-1}(t)\,dt + \mathbf{x}(0)$$

With the initial approximation $\mathbf{x}_0(t) = \mathbf{x}(0)$, obtain the nth approximation to the solution. Show that as n is increased to infinity the solution approaches $e^{\mathbf{A}t}\mathbf{x}(0)$.

Solution The successive approximations are

$$\mathbf{x}_0(t) = \mathbf{x}(0)$$
$$\mathbf{x}_1(t) = (\mathbf{I} + \mathbf{A}t)\mathbf{x}(0)$$
$$\mathbf{x}_2(t) = \left(\mathbf{I} + \mathbf{A}t + \frac{1}{2!}\mathbf{A}^2t^2\right)\mathbf{x}(0)$$
$$\cdots \qquad \cdots \qquad \cdots$$
$$\mathbf{x}_n(t) = \left(\mathbf{I} + \mathbf{A}t + \frac{1}{2!}\mathbf{A}^2t^2 + \cdots + \frac{1}{n!}\mathbf{A}^nt^n\right)\mathbf{x}(0)$$

As $n \longrightarrow \infty$

$$\begin{aligned} \mathbf{x}(t) &= \lim_{n\to\infty} \mathbf{x}_n(t) \\ &= \lim_{n\to\infty} \left(\mathbf{I} + \mathbf{A}t + \frac{1}{2!}\mathbf{A}^2t^2 + \cdots + \frac{1}{n!}\mathbf{A}^nt^n\right)\mathbf{x}(0) \\ &= \sum_{n=0}^{\infty} \frac{1}{n!}\mathbf{A}^nt^n\mathbf{x}(0) \\ &= e^{\mathbf{A}t}\mathbf{x}(0) \end{aligned}$$

Problem A-6-10 By use of Method 4, Section 6–4, compute $e^{\mathbf{A}t}$, where

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 4 \\ 0 & 2 & 0 \\ 0 & 3 & 1 \end{bmatrix}$$

Solution The eigenvalues of $\mathbf{A}$ are 2, 2, 1. The characteristic polynomial and the minimal polynomial are the same for this $\mathbf{A}$. The minimal polynomial is given by

$$\phi(\lambda) = (2 - \lambda)^2(1 - \lambda)$$

In this example, $e^{\mathbf{A}t}$ can be written as

$$e^{\mathbf{A}t} = \alpha_0\mathbf{I} + \alpha_1\mathbf{A} + \alpha_2\mathbf{A}^2$$

Equations for the determination of α_0, α_1, and α_2 are

$$\begin{aligned} \alpha_1 + 2\alpha_2\lambda_1 &= te^{\lambda_1 t} \\ \alpha_0 + \alpha_1\lambda_1 + \alpha_2\lambda_1^2 &= e^{\lambda_1 t} \\ \alpha_0 + \alpha_1\lambda_3 + \alpha_2\lambda_3^2 &= e^{\lambda_3 t} \end{aligned} \tag{6-76}$$

Substituting $\lambda_1 = 2, \lambda_3 = 1$ into Eq. (6–76),

$$\begin{aligned} \alpha_1 + 4\alpha_2 &= te^{2t} \\ \alpha_0 + 2\alpha_1 + 4\alpha_2 &= e^{2t} \\ \alpha_0 + \alpha_1 + \alpha_2 &= e^{t} \end{aligned}$$

Solving for α_0, α_1, and α_2 we obtain

$$\begin{aligned} \alpha_0 &= 4e^t - 3e^{2t} + 2te^{2t} \\ \alpha_1 &= -4e^t + 4e^{2t} - 3te^{2t} \\ \alpha_2 &= e^t - e^{2t} + te^{2t} \end{aligned}$$

Hence,

$$\begin{aligned} e^{\mathbf{A}t} &= (4e^t - 3e^{2t} + 2te^{2t})\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &\quad + (-4e^t + 4e^{2t} - 3te^{2t})\begin{bmatrix} 2 & 1 & 4 \\ 0 & 2 & 0 \\ 0 & 3 & 1 \end{bmatrix} \\ &\quad + (e^t - e^{2t} + te^{2t})\begin{bmatrix} 4 & 16 & 12 \\ 0 & 4 & 0 \\ 0 & 9 & 1 \end{bmatrix} \\ &= \begin{bmatrix} e^{2t} & 12e^t - 12e^{2t} + 13te^{2t} & -4e^t + 4e^{2t} \\ 0 & e^{2t} & 0 \\ 0 & -3e^t + 3e^{2t} & e^t \end{bmatrix} \end{aligned}$$

Problem A-6-11 Consider the system

$$\overset{(n)}{x} + a_1(t)\overset{(n-1)}{x} + \cdots + a_n(t)x = 0$$

Define

$$W(t) = \begin{vmatrix} \phi_1 & \phi_2 & \cdots & \phi_n \\ \dot{\phi}_1 & \dot{\phi}_2 & \cdots & \dot{\phi}_n \\ \vdots & \vdots & & \vdots \\ \overset{(n-1)}{\phi_1} & \overset{(n-1)}{\phi_2} & \cdots & \overset{(n-1)}{\phi_n} \end{vmatrix}$$

where $\phi_1, \phi_2, \ldots, \phi_n$ are n independent solutions of the given system. Show that

$$\dot{W}(t) + a_1(t)W(t) = 0 \qquad \text{or} \qquad W(t) = W(t_0)e^{-\int_{t_0}^{t} a_1(\tau)\,d\tau}$$

Solution Let us differentiate $W(t)$ with respect to t.

$$\dot{W}(t) = \begin{vmatrix} \dot{\phi}_1 & \dot{\phi}_2 & \cdots & \dot{\phi}_n \\ \dot{\phi}_1 & \dot{\phi}_2 & \cdots & \dot{\phi}_n \\ \vdots & \vdots & & \vdots \\ \overset{(n-1)}{\phi_1} & \overset{(n-1)}{\phi_2} & \cdots & \overset{(n-1)}{\phi_n} \end{vmatrix} + \begin{vmatrix} \phi_1 & \phi_2 & \cdots & \phi_n \\ \ddot{\phi}_1 & \ddot{\phi}_2 & \cdots & \ddot{\phi}_n \\ \vdots & \vdots & & \vdots \\ \overset{(n-1)}{\phi_1} & \overset{(n-1)}{\phi_2} & \cdots & \overset{(n-1)}{\phi_n} \end{vmatrix}$$

$$+ \cdots + \begin{vmatrix} \phi_1 & \phi_2 & \cdots & \phi_n \\ \dot{\phi}_1 & \dot{\phi}_2 & \cdots & \dot{\phi}_n \\ \vdots & \vdots & & \vdots \\ \overset{(n-2)}{\phi_1} & \overset{(n-2)}{\phi_2} & \cdots & \overset{(n-2)}{\phi_n} \\ \overset{(n)}{\phi_1} & \overset{(n)}{\phi_2} & \cdots & \overset{(n)}{\phi_n} \end{vmatrix}$$

$$= \begin{vmatrix} \phi_1 & \phi_2 & \cdots & \phi_n \\ \dot{\phi}_1 & \dot{\phi}_2 & \cdots & \dot{\phi}_n \\ \vdots & \vdots & & \vdots \\ \overset{(n-2)}{\phi_1} & \overset{(n-2)}{\phi_2} & \cdots & \overset{(n-2)}{\phi_n} \\ \overset{(n)}{\phi_1} & \overset{(n)}{\phi_2} & \cdots & \overset{(n)}{\phi_n} \end{vmatrix}$$

Notice that

$$\overset{(n)}{\phi_i} = -a_1(t)\overset{(n-1)}{\phi_i} - a_2(t)\overset{(n-2)}{\phi_i} - \cdots - a_n(t)\phi_i$$

Hence we obtain

$$\dot{W}(t) = \begin{vmatrix} \phi_1 & \phi_2 & \cdots & \phi_n \\ \dot{\phi}_1 & \dot{\phi}_2 & \cdots & \dot{\phi}_n \\ \vdots & \vdots & & \vdots \\ \overset{(n-2)}{\phi_1} & \overset{(n-2)}{\phi_2} & \cdots & \overset{(n-2)}{\phi_n} \\ -a_1(t)\overset{(n-1)}{\phi_1} & -a_1(t)\overset{(n-1)}{\phi_2} & \cdots & -a_1(t)\overset{(n-1)}{\phi_n} \end{vmatrix}$$

$$= -a_1(t)W(t)$$

Integrating,

$$W(t) = W(t_0)e^{-\int_{t_0}^{t} a_1(\tau)\,d\tau}$$

Problem A-6-12 (6–9) Compute $e^{\mathbf{A}t}$ where

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

by use of Sylvester's interpolation formula.

Solution We must first find the roots of the following characteristic equation:

$$|\mathbf{A} - \lambda\mathbf{I}| = \begin{vmatrix} -\lambda & 0 & 1 & 0 \\ 1 & -\lambda & 0 & 1 \\ 0 & 0 & -\lambda & -1 \\ 0 & 1 & 0 & -\lambda \end{vmatrix} = \lambda^4 - \lambda^2 + 1 = 0$$

The roots are

$$\frac{-\sqrt{3}+j}{2}, \quad \frac{\sqrt{3}-j}{2}, \quad \frac{-\sqrt{3}-j}{2}, \quad \frac{\sqrt{3}+j}{2}$$

Let us define

$$\frac{-\sqrt{3}+j}{2} = e^{j(5/6)\pi} = \lambda_1$$

$$\frac{\sqrt{3}-j}{2} = -e^{j(5/6)\pi} = \lambda_2$$

$$\frac{-\sqrt{3}-j}{2} = e^{-j(5/6)\pi} = \lambda_3$$

$$\frac{\sqrt{3}+j}{2} = -e^{-j(5/6)\pi} = \lambda_4$$

Sylvester's interpolation formula can be written as follows:

$$\begin{vmatrix} 1 & \lambda_1 & \lambda_1^2 & \lambda_1^3 & e^{\lambda_1 t} \\ 1 & \lambda_2 & \lambda_2^2 & \lambda_2^3 & e^{\lambda_2 t} \\ 1 & \lambda_3 & \lambda_3^2 & \lambda_3^3 & e^{\lambda_3 t} \\ 1 & \lambda_4 & \lambda_4^2 & \lambda_4^3 & e^{\lambda_4 t} \\ \mathbf{I} & \mathbf{A} & \mathbf{A}^2 & \mathbf{A}^3 & e^{\mathbf{A}t} \end{vmatrix} = 0$$

Expanding the determinant about the last column we obtain

$$\begin{vmatrix} 1 & \lambda_1 & \lambda_1^2 & \lambda_1^3 \\ 1 & \lambda_2 & \lambda_2^2 & \lambda_2^3 \\ 1 & \lambda_3 & \lambda_3^2 & \lambda_3^3 \\ 1 & \lambda_4 & \lambda_4^2 & \lambda_4^3 \end{vmatrix} e^{\mathbf{A}t} = -\begin{vmatrix} 1 & \lambda_2 & \lambda_2^2 & \lambda_2^3 \\ 1 & \lambda_3 & \lambda_3^2 & \lambda_3^3 \\ 1 & \lambda_4 & \lambda_4^2 & \lambda_4^3 \\ \mathbf{I} & \mathbf{A} & \mathbf{A}^2 & \mathbf{A}^3 \end{vmatrix} e^{\lambda_1 t} + \begin{vmatrix} 1 & \lambda_1 & \lambda_1^2 & \lambda_1^3 \\ 1 & \lambda_3 & \lambda_3^2 & \lambda_3^3 \\ 1 & \lambda_4 & \lambda_4^2 & \lambda_4^3 \\ \mathbf{I} & \mathbf{A} & \mathbf{A}^2 & \mathbf{A}^3 \end{vmatrix} e^{\lambda_2 t}$$

$$- \begin{vmatrix} 1 & \lambda_1 & \lambda_1^2 & \lambda_1^3 \\ 1 & \lambda_2 & \lambda_2^2 & \lambda_2^3 \\ 1 & \lambda_4 & \lambda_4^2 & \lambda_4^3 \\ \mathbf{I} & \mathbf{A} & \mathbf{A}^2 & \mathbf{A}^3 \end{vmatrix} e^{\lambda_3 t} + \begin{vmatrix} 1 & \lambda_1 & \lambda_1^2 & \lambda_1^3 \\ 1 & \lambda_2 & \lambda_2^2 & \lambda_2^3 \\ 1 & \lambda_3 & \lambda_3^2 & \lambda_3^3 \\ \mathbf{I} & \mathbf{A} & \mathbf{A}^2 & \mathbf{A}^3 \end{vmatrix} e^{\lambda_4 t}$$

The preceding 4×4 determinants may be evaluated as follows:

$$\begin{vmatrix} 1 & \lambda_1 & \lambda_1^2 & \lambda_1^3 \\ 1 & \lambda_2 & \lambda_2^2 & \lambda_2^3 \\ 1 & \lambda_3 & \lambda_3^2 & \lambda_3^3 \\ 1 & \lambda_4 & \lambda_4^2 & \lambda_4^3 \end{vmatrix} = (\lambda_4 - \lambda_3)(\lambda_4 - \lambda_2)(\lambda_4 - \lambda_1) \cdot (\lambda_3 - \lambda_2)(\lambda_3 - \lambda_1)(\lambda_2 - \lambda_1)$$

$$= -12$$

$$\begin{vmatrix} 1 & \lambda_2 & \lambda_2^2 & \lambda_2^3 \\ 1 & \lambda_3 & \lambda_3^2 & \lambda_3^3 \\ 1 & \lambda_4 & \lambda_4^2 & \lambda_4^3 \\ \mathbf{I} & \mathbf{A} & \mathbf{A}^2 & \mathbf{A}^3 \end{vmatrix} = (\mathbf{A} - \lambda_4\mathbf{I})(\mathbf{A} - \lambda_3\mathbf{I})(\mathbf{A} - \lambda_2\mathbf{I}) \cdot (\lambda_4 - \lambda_3)(\lambda_4 - \lambda_2)(\lambda_3 - \lambda_2)$$

$$= (3 - j\sqrt{3})\mathbf{I} - 2\sqrt{3}\,\mathbf{A} + j2\sqrt{3}\,\mathbf{A}^2 + (\sqrt{3} - j3)\mathbf{A}^3$$

Similarly, we can evaluate the other determinants. Then $e^{\mathbf{A}t}$ can be written as follows:

$$\begin{aligned} e^{\mathbf{A}t} = -\tfrac{1}{12}\{&[-(3 - j\sqrt{3})\mathbf{I} + 2\sqrt{3}\,\mathbf{A} - j2\sqrt{3}\,\mathbf{A}^2 - (\sqrt{3} - j3)\mathbf{A}^3]e^{\lambda_1 t} \\ &+ [-(3 - j\sqrt{3})\mathbf{I} - 2\sqrt{3}\,\mathbf{A} - j2\sqrt{3}\,\mathbf{A}^2 + (\sqrt{3} - j3)\mathbf{A}^3]e^{\lambda_2 t} \\ &+ [-(3 + j\sqrt{3})\mathbf{I} + 2\sqrt{3}\,\mathbf{A} + j2\sqrt{3}\,\mathbf{A}^2 - (\sqrt{3} + j3)\mathbf{A}^3]e^{\lambda_3 t} \\ &+ [-(3 + j\sqrt{3})\mathbf{I} - 2\sqrt{3}\,\mathbf{A} + j2\sqrt{3}\,\mathbf{A}^2 + (\sqrt{3} + j3)\mathbf{A}^3]e^{\lambda_4 t}\} \end{aligned}$$

Grouping real and imaginary terms and using the identity

$$e^{j\theta} = \cos\theta + j\sin\theta$$

$e^{\mathbf{A}t}$ becomes

$$\begin{aligned} e^{\mathbf{A}t} = \frac{\sqrt{3}}{6}\Big\{&e^{(\sqrt{3}/2)t}\left[(\sqrt{3}\,\mathbf{I} + 2\mathbf{A} - \mathbf{A}^3)\cos\frac{t}{2} + (-\mathbf{I} + 2\mathbf{A}^2 + \sqrt{3}\,\mathbf{A}^3)\sin\frac{t}{2}\right] \\ &+ e^{-(\sqrt{3}/2)t}\left[(\sqrt{3}\,\mathbf{I} - 2\mathbf{A} + \mathbf{A}^3)\cos\frac{t}{2} + (\mathbf{I} - 2\mathbf{A}^2 + \sqrt{3}\,\mathbf{A}^3)\sin\frac{t}{2}\right]\Big\} \end{aligned}$$

Denoting

$$e^{(\sqrt{3}/2)t} + e^{-(\sqrt{3}/2)t} = k$$

$$e^{(\sqrt{3}/2)t} - e^{-(\sqrt{3}/2)t} = h$$

$e^{\mathbf{A}t}$ can be written as

$$\begin{aligned} e^{\mathbf{A}t} = \frac{\sqrt{3}}{6}\Big[&\mathbf{I}\left(\sqrt{3}\,k\cos\frac{t}{2} - h\sin\frac{t}{2}\right) + \mathbf{A}2h\cos\frac{t}{2} + \mathbf{A}^2 2h\sin\frac{t}{2} \\ &+ \mathbf{A}^3\left(-h\cos\frac{t}{2} + \sqrt{3}\,k\sin\frac{t}{2}\right)\Big] \end{aligned}$$

Since $\mathbf{A}^2$ and $\mathbf{A}^3$ are obtained as

$$\mathbf{A}^2 = \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \qquad \mathbf{A}^3 = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

$e^{\mathbf{A}t}$ has the form shown in the following:

$$e^{\mathbf{A}t} = \begin{bmatrix} \frac{k}{2}\cos\frac{t}{2} - \frac{h}{2\sqrt{3}}\sin\frac{t}{2} & \frac{h}{2\sqrt{3}}\cos\frac{t}{2} - \frac{k}{2}\sin\frac{t}{2} \\ \frac{h}{2\sqrt{3}}\cos\frac{t}{2} + \frac{k}{2}\sin\frac{t}{2} & \frac{k}{2}\cos\frac{t}{2} + \frac{h}{2\sqrt{3}}\sin\frac{t}{2} \\ \frac{h}{2\sqrt{3}}\cos\frac{t}{2} - \frac{k}{2}\sin\frac{t}{2} & -\frac{h}{\sqrt{3}}\sin\frac{t}{2} \\ \frac{h}{\sqrt{3}}\sin\frac{t}{2} & \frac{h}{2\sqrt{3}}\cos\frac{t}{2} + \frac{k}{2}\sin\frac{t}{2} \end{bmatrix}$$

$$\begin{bmatrix} \frac{h}{\sqrt{3}}\cos\frac{t}{2} & -\frac{h}{\sqrt{3}}\sin\frac{t}{2} \\ \frac{h}{\sqrt{3}}\sin\frac{t}{2} & \frac{h}{\sqrt{3}}\cos\frac{t}{2} \\ \frac{k}{2}\cos\frac{t}{2} - \frac{h}{2\sqrt{3}}\sin\frac{t}{2} & -\frac{h}{2\sqrt{3}}\cos\frac{t}{2} - \frac{k}{2}\sin\frac{t}{2} \\ -\frac{h}{2\sqrt{3}}\cos\frac{t}{2} + \frac{k}{2}\sin\frac{t}{2} & \frac{k}{2}\cos\frac{t}{2} + \frac{h}{2\sqrt{3}}\sin\frac{t}{2} \end{bmatrix}$$

By letting

$$k = 2p$$
$$h = 2\sqrt{3}\,q$$
$$\sin\frac{t}{2} = a$$
$$\cos\frac{t}{2} = b$$

$e^{\mathbf{A}t}$ can be simplified into

$$e^{\mathbf{A}t} = \begin{bmatrix} pb - qa & -pa + qb & 2qb & -2qa \\ pa + qb & pb + qa & 2qa & 2qb \\ -pa + qb & -2qa & pb - qa & -pa - qb \\ 2qa & pa + qb & pa - qb & pb + qa \end{bmatrix}$$

Problem A-6-13 Obtain the solution of

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

by use of the Laplace transform method.

Solution Let us define

$$\mathbf{A} = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}$$

Then

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}(0)$$

where $e^{\mathbf{A}t}$ is the inverse Laplace transform of $(s\mathbf{I} - \mathbf{A})^{-1}$ which is evaluated as

$$(s\mathbf{I} - \mathbf{A})^{-1} = \begin{bmatrix} s+1 & -1 \\ -1 & s-1 \end{bmatrix}^{-1} = \frac{1}{s^2 - 2}\begin{bmatrix} s-1 & 1 \\ 1 & s+1 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{s-1}{(s+\sqrt{2})(s-\sqrt{2})} & \frac{1}{(s+\sqrt{2})(s-\sqrt{2})} \\ \frac{1}{(s+\sqrt{2})(s-\sqrt{2})} & \frac{s+1}{(s+\sqrt{2})(s-\sqrt{2})} \end{bmatrix}$$

Taking the inverse Laplace transform of $(s\mathbf{I} - \mathbf{A})^{-1}$, we obtain the solution $\mathbf{x}(t)$ as follows:

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\sqrt{2}+1}{2\sqrt{2}}e^{-\sqrt{2}t} + \frac{\sqrt{2}-1}{2\sqrt{2}}e^{\sqrt{2}t} & -\frac{1}{2\sqrt{2}}e^{-\sqrt{2}t} + \frac{1}{2\sqrt{2}}e^{\sqrt{2}t} \\ \frac{-1}{2\sqrt{2}}e^{-\sqrt{2}t} + \frac{1}{2\sqrt{2}}e^{\sqrt{2}t} & \frac{\sqrt{2}-1}{2\sqrt{2}}e^{-\sqrt{2}t} + \frac{\sqrt{2}+1}{2\sqrt{2}}e^{\sqrt{2}t} \end{bmatrix} \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix}$$

Problem A-6-14[1] Draw an analog computer diagram for solving the following simultaneous linear equations:

$$a_{11}x_1 + a_{12}x_2 = b_1$$
$$a_{21}x_1 + a_{22}x_2 = b_2$$

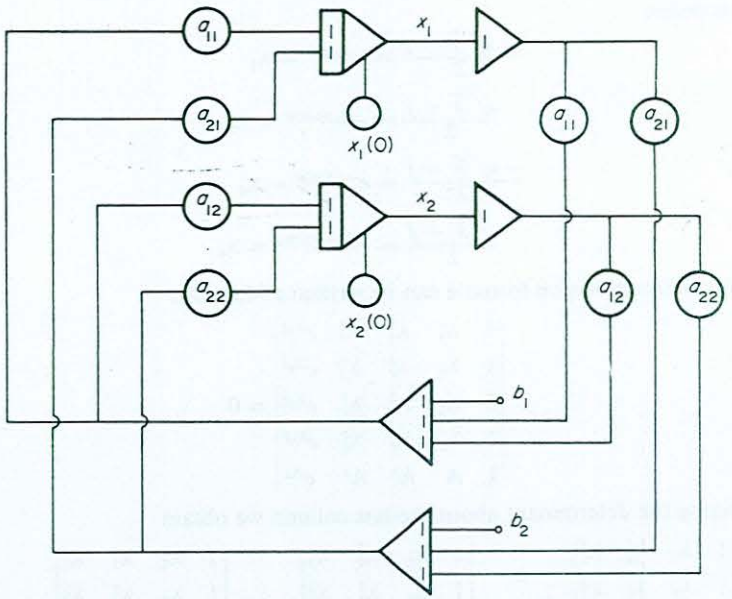


Figure 6-5 Analog computer diagram for solving $\mathbf{Ax} = \mathbf{B}$.

[1]A technique similar to the one presented here can be used for finding the eigenvalues and eigenvectors of an $n \times n$ matrix $\mathbf{A}$. For detail see (*6–10*).

Assume that $0 \leqslant a_{ij} \leqslant 1$, the b_i are real, and

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0$$

Solution The solution of the given equation

$$\mathbf{Ax} = \mathbf{B}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

can be obtained as a steady state solution of the following equation:

$$\dot{\mathbf{x}} + \mathbf{A}'\mathbf{Ax} = \mathbf{A}'\mathbf{B}$$

This vector matrix differential equation is stable and possesses a steady state solution, $\mathbf{x} = \mathbf{A}^{-1}\mathbf{B}$. Figure 6–5 shows a possible analog computer diagram for the present problem. The diagram can be easily modified to solve n simultaneous equations with any real coefficients a_{ij} and any real b_i.

Problem A-6-15 Find the solution of

$$\frac{d}{dt}\mathbf{B}(t) = \mathbf{A}(t)\mathbf{B}(t) + \mathbf{B}(t)\mathbf{A}^*(t), \qquad \mathbf{B}(t_0) = \mathbf{B}_0$$

where the elements of $\mathbf{A}(t)$ are continuous functions of time.

Solution The solution is given by

$$\mathbf{B}(t) = \Phi(t, t_0)\mathbf{B}_0\Phi^*(t, t_0)$$

where $\Phi(t, t_0)$ is the solution of the following matrix differential equation:

$$\frac{d}{dt}\Phi(t, t_0) = \mathbf{A}(t)\Phi(t, t_0), \qquad \Phi(t_0, t_0) = \mathbf{I}$$

To verify that $\Phi(t, t_0)\mathbf{B}_0\Phi^*(t, t_0)$ is the solution, use direct substitution. Since the elements of $\mathbf{A}(t)$ are continuous, the solution is unique.

Problem A-6-16 Show that the solution of the following matrix differential equation:

$$\frac{d\mathbf{X}}{dt} = \mathbf{AX} + \mathbf{XB}, \qquad \mathbf{X}(0) = \mathbf{C}$$

is given by

$$\mathbf{X} = e^{\mathbf{A}t}\mathbf{C}\,e^{\mathbf{B}t}$$

Solution Differentiate $\mathbf{X}$ with respect to t. Then we obtain

$$\frac{d\mathbf{X}}{dt} = \mathbf{A}\,e^{\mathbf{A}t}\mathbf{C}\,e^{\mathbf{B}t} + e^{\mathbf{A}t}\mathbf{C}\mathbf{B}\,e^{\mathbf{B}t}$$

Since $\mathbf{B}$ and $e^{\mathbf{B}t}$ commute, we have

$$e^{\mathbf{A}t}\mathbf{C}\mathbf{B}\,e^{\mathbf{B}t} = e^{\mathbf{A}t}\mathbf{C}\,e^{\mathbf{B}t}\mathbf{B}$$

Hence,

$$\frac{d\mathbf{X}}{dt} = \mathbf{AX} + \mathbf{XB}$$

or $\mathbf{X} = e^{\mathbf{A}t}\mathbf{C}\,e^{\mathbf{B}t}$ is the solution of the given matrix differential equation with the given initial condition.

Problem A-6-17 Consider the following equation:

$$\mathbf{A}'\mathbf{B} + \mathbf{BA} = -\mathbf{C} \tag{6–77}$$

where $\mathbf{A}$ is a stable matrix (meaning that the eigenvalues of $\mathbf{A}$ have negative real parts) and $\mathbf{B}$ and $\mathbf{C}$ are symmetric matrices. Show that the matrix $\mathbf{B}$ satisfying Eq. (6–77) can be given by

$$\mathbf{B} = \int_0^\infty e^{\mathbf{A}'t}\mathbf{C}\,e^{\mathbf{A}t}\,dt$$

Solution Consider the matrix differential equation

$$\dot{\mathbf{X}} = \mathbf{A}'\mathbf{X} + \mathbf{XA}, \qquad \mathbf{X}(0) = \mathbf{C} \tag{6–78}$$

The solution of Eq. (6–78) is given by

$$\mathbf{X} = e^{\mathbf{A}'t}\mathbf{C}\,e^{\mathbf{A}t}$$

Integrating both sides of Eq. (6–78) from $t = 0$ to $t = \infty$, we obtain

$$\int_0^\infty \dot{\mathbf{X}}\,dt = \int_0^\infty \mathbf{A}'\mathbf{X}\,dt + \int_0^\infty \mathbf{XA}\,dt$$

or

$$\mathbf{X}(\infty) - \mathbf{X}(0) = \mathbf{A}'\left(\int_0^\infty \mathbf{X}\,dt\right) + \left(\int_0^\infty \mathbf{X}\,dt\right)\mathbf{A}$$

Since $\mathbf{A}$ is a stable matrix, $\mathbf{X}(\infty) = \mathbf{0}$. Therefore, we have

$$-\mathbf{C} = \mathbf{A}'\left(\int_0^\infty \mathbf{X}\,dt\right) + \left(\int_0^\infty \mathbf{X}\,dt\right)\mathbf{A}$$

Hence,

$$\mathbf{B} = \int_0^\infty \mathbf{X}\,dt = \int_0^\infty e^{\mathbf{A}'t}\mathbf{C}\,e^{\mathbf{A}t}\,dt$$

Problem A-6-18 Consider the following system

$$\overset{(n)}{x} + a_1\overset{(n-1)}{x} + \cdots + a_{n-1}\dot{x} + a_n x = \dot{u} + bu$$

where

$$x(0) = \dot{x}(0) = \cdots = \overset{(n-1)}{x}(0) = 0$$

Show that if the forcing function $u(t)$ is given by

$$\begin{aligned} u(t) &= 0, & t < 0 \\ &= e^{-bt}, & t \geqslant 0 \end{aligned}$$

then the response $x(t)$ is identical with the impulse response function $h(t)$, where $h(t)$ satisfies

$$\overset{(n)}{h} + a_1\overset{(n-1)}{h} + \cdots + a_{n-1}\dot{h} + a_n h = \delta(t)$$

Solution For the given $u(t)$ we obtain

$$\dot{u}(t) = \delta(t)e^{-bt} - b\,e^{-bt}$$

Hence for $t \geqslant 0$

$$\begin{aligned} \dot{u}(t) + bu(t) &= \delta(t)e^{-bt} - b\,e^{-bt} + b\,e^{-bt} \\ &= \delta(t)e^{-bt} \\ &= \delta(t) \end{aligned}$$

Thus, the response of the system is given by the impulse response function $h(t)$.

Problem A-6-19 Obtain the response of the following system:

$$\frac{X(s)}{U(s)} = \frac{2(s+3)}{(s+1)(s+2)}, \qquad x(0) = \dot{x}(0) = 0$$

where

$$u(t) = 0, \qquad t < 0$$
$$= e^{-t}, \qquad t \geqslant 0$$

(1) Use the Laplace transform approach.
(2) Use the partial-fraction expansion technique.

Solution

(1) The Laplace transform of $u(t)$ is

$$U(s) = \frac{1}{s+1}$$

Hence

$$X(s) = \frac{2(s+3)}{(s+1)^2(s+2)} = \frac{-2}{s+1} + \frac{4}{(s+1)^2} + \frac{2}{s+2}$$

The inverse Laplace transform of $X(s)$ yields

$$x(t) = -2e^{-t} + 4te^{-t} + 2e^{-2t}$$

(2)

$$\frac{X(s)}{U(s)} = \frac{X_1(s)}{U(s)} = \frac{4}{s+1} + \frac{-2}{s+2}$$

$$\frac{sX(s) - 2U(s)}{U(s)} = \frac{X_2(s)}{U(s)} = \frac{-4}{s+1} + \frac{4}{s+2}$$

Hence,

$$\begin{bmatrix} \dfrac{X_1(s)}{U(s)} \\ \dfrac{X_2(s)}{U(s)} \end{bmatrix} = \begin{bmatrix} 4 & -2 \\ -4 & 4 \end{bmatrix} \begin{bmatrix} \dfrac{1}{s+1} \\ \dfrac{1}{s+2} \end{bmatrix}$$

Let

$$\begin{bmatrix} X_1(s) \\ X_2(s) \end{bmatrix} = \begin{bmatrix} 4 & -2 \\ -4 & 4 \end{bmatrix} \begin{bmatrix} Y_1(s) \\ Y_2(s) \end{bmatrix}$$

Then

$$\begin{bmatrix} \dfrac{Y_1(s)}{U(s)} \\ \dfrac{Y_2(s)}{U(s)} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{s+1} \\ \dfrac{1}{s+2} \end{bmatrix}$$

and $y_1(t)$ and $y_2(t)$ are obtained as

$$y_1(t) = \mathscr{L}^{-1}[Y_1(s)] = \mathscr{L}^{-1}\left[\frac{1}{(s+1)^2}\right] = te^{-t}$$

$$y_2(t) = \mathscr{L}^{-1}[Y_2(s)] = \mathscr{L}^{-1}\left[\frac{1}{(s+1)(s+2)}\right] = e^{-t} - e^{-2t}$$

Then $x_1(t)$ and $x_2(t)$ are found as

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} 4y_1(t) - 2y_2(t) \\ -4y_1(t) + 4y_2(t) \end{bmatrix} = \begin{bmatrix} 4te^{-t} - 2e^{-t} + 2e^{-2t} \\ -4te^{-t} + 4e^{-t} - 4e^{-2t} \end{bmatrix}$$

Notice that the given transfer function can be written in the time domain as follows:

$$\ddot{x} + 3\dot{x} + 2x = 2\dot{u} + 6u$$

A state space representation of this equation can be given as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 2 \\ 0 \end{bmatrix} [u]$$

where

$$x_1 = x$$
$$x_2 = \dot{x}_1 - 2u$$

Problem A-6-20 Obtain the response of the following system:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 2 \\ 0 \end{bmatrix} [u], \qquad \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

where

$$u(t) = 0, \qquad t < 0$$
$$= e^{-t}, \qquad t \geqslant 0$$

Solution Let $\mathbf{x} = \mathbf{P}\mathbf{y}$ where

$$\mathbf{P} = \begin{bmatrix} 1 & 1 \\ -1 & -2 \end{bmatrix}$$

Then we obtain

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} 2 & 1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \end{bmatrix} [u]$$
$$= \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} 4 \\ -2 \end{bmatrix} [u] \qquad (6\text{–}79)$$

or

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y} + \mathbf{B}u$$

The solution of Eq. (6–79) is given by

$$\mathbf{y}(t) = e^{\mathbf{A}t}\left[\mathbf{y}(0) + \int_0^t e^{-\mathbf{A}\tau}\mathbf{B}u(\tau)d\tau\right]$$

where

$$e^{\mathbf{A}t} = \begin{bmatrix} e^{-t} & 0 \\ 0 & e^{-2t} \end{bmatrix}$$

Thus we obtain

$$\begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} e^{-t} & 0 \\ 0 & e^{-2t} \end{bmatrix} \left\{ \begin{bmatrix} y_1(0) \\ y_2(0) \end{bmatrix} + \int_0^t \begin{bmatrix} e^{\tau} & 0 \\ 0 & e^{2\tau} \end{bmatrix} \begin{bmatrix} 4 \\ -2 \end{bmatrix} e^{-\tau} d\tau \right\}$$
$$= \begin{bmatrix} e^{-t} & 0 \\ 0 & e^{-2t} \end{bmatrix} \begin{bmatrix} y_1(0) \\ y_2(0) \end{bmatrix} + \begin{bmatrix} e^{-t} & 0 \\ 0 & e^{-2t} \end{bmatrix} \begin{bmatrix} 4t \\ -2(e^t - 1) \end{bmatrix}$$

Since

$$\mathbf{y}(0) = \mathbf{P}^{-1}\mathbf{x}(0) = \mathbf{0}$$

we obtain

$$\begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} e^{-t} & 0 \\ 0 & e^{-2t} \end{bmatrix} \begin{bmatrix} 4t \\ -2(e^t - 1) \end{bmatrix} = \begin{bmatrix} 4te^{-t} \\ -2e^{-t} + 2e^{-2t} \end{bmatrix}$$

Since

$$\mathbf{x}(t) = \mathbf{P}\mathbf{y}(t) = \begin{bmatrix} 1 & 1 \\ -1 & -2 \end{bmatrix} \begin{bmatrix} 4te^{-t} \\ -2e^{-t} + 2e^{-2t} \end{bmatrix}$$

we obtain

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} 4te^{-t} - 2e^{-t} + 2e^{-2t} \\ -4te^{-t} + 4e^{-t} - 4e^{-2t} \end{bmatrix}$$

Problem A-6-21 Obtain the response of the following time-varying system:

$$\dot{x} + tx = u$$

where u is an arbitrary input.

Solution The solution of this system is given by either

$$x(t) = \phi(t, t_0)x(t_0) + \phi(t, t_0) \int_{t_0}^{t} \phi^{-1}(\tau, t_0)u(\tau)d\tau$$

or

$$x(t) = \phi(t, t_0)x(t_0) + \int_{t_0}^{t} h(t, \tau)u(\tau)d\tau$$

Notice that

$$\phi(t, t_0)\phi^{-1}(\tau, t_0) = h(t, \tau)$$

In this system since t and $\int_{t_0}^{t} \tau d\tau$ commute for all t, $\phi(t, t_0)$ can be expressed as

$$\phi(t, t_0) = e^{-\int_{t_0}^{t} \tau\, d\tau} = e^{-[(t^2/2) - (t_0^2/2)]}$$

Hence

$$\phi^{-1}(\tau, t_0) = e^{(\tau^2/2) - (t_0^2/2)}$$

Therefore,

$$h(t, \tau) = e^{-[(t^2/2) - (\tau^2/2)]}, \qquad t \geqslant \tau$$
$$= 0 \qquad t < \tau$$

The output $x(t)$ is thus given by

$$x(t) = e^{-[(t^2/2) - (t_0^2/2)]}x(t_0) + e^{-t^2/2} \int_{t_0}^{t} e^{\tau^2/2}u(\tau)d\tau$$

Problem A-6-22 Show that

$$\ddot{x} + a(t)\dot{x} + b(t)x = 0$$

can be reduced to the following form:

$$\ddot{z} + c(t)z = 0$$

Solution Let

$$x = ze^{-(1/2)\int_{t_0}^{t} a(t)dt} = zy$$

Then,

$$\dot{x} = y\dot{z} - \tfrac{1}{2}a(t)yz$$
$$\ddot{x} = y\ddot{z} - a(t)y\dot{z} - \tfrac{1}{2}\dot{a}(t)yz + \tfrac{1}{4}a^2(t)yz$$

Hence,

$$\ddot{x} + a(t)\dot{x} + b(t)x$$
$$= y\{\ddot{z} + [b(t) - \tfrac{1}{2}\dot{a}(t) - \tfrac{1}{4}a^2(t)]z\}$$

The equivalent equation is, therefore,

$$\ddot{z} + c(t)z = 0$$

where

$$c(t) = b(t) - \tfrac{1}{2}\dot{a}(t) - \tfrac{1}{4}a^2(t)$$

Problem A-6-23 Show that the solution of the following scalar differential equation:

$$\ddot{x} + a(t)\dot{x} + b(t)x = u \qquad (6\text{–}80)$$
$$x(t_0) = \dot{x}(t_0) = 0$$

can be given by

$$x(t) = x_1(t) = -x_{11}(t) \int_{t_0}^{t} \frac{x_{12}(\tau)u(\tau)}{W(x_{11}, x_{12})}\, d\tau$$
$$+x_{12}(t) \int_{t_0}^{t} \frac{x_{11}(\tau)u(\tau)}{W(x_{11}, x_{12})}\, d\tau$$

where $x_{11}(t)$ and $x_{12}(t)$ are linearly independent solutions of the homogeneous equation

$$\ddot{x} + a(t)\dot{x} + b(t)x = 0$$

and W is the Wronskian given by

$$W(x_{11}, x_{12}) = \begin{vmatrix} x_{11}(t) & x_{12}(t) \\ \dot{x}_{11}(t) & \dot{x}_{12}(t) \end{vmatrix} \neq 0$$

Solution Equation (6–80) can be written as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -b(t) & -a(t) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} [u] \qquad (6\text{–}81)$$

where

$$x_1 = x, \qquad x_2 = \dot{x}$$

The solution of Eq. (6–81) is given by

$$\mathbf{x}(t) = \Phi(t, t_0)\mathbf{x}_0 + \Phi(t, t_0) \int_{t_0}^{t} \Phi^{-1}(\tau, t_0)\mathbf{B}u\, d\tau$$

where $\Phi(t, t_0)$ is the fundamental matrix satisfying the condition that $\Phi(t_0, t_0) = \mathbf{I}$, and $\mathbf{B}$ is given by

$$\mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Since $x(t_0) = \dot{x}(t_0) = 0$, we have $\mathbf{x}_0 = \mathbf{0}$; $\mathbf{x}(t)$ is, therefore, given by

$$\mathbf{x}(t) = \Phi(t, t_0) \int_{t_0}^{t} \Phi^{-1}(\tau, t_0)\mathbf{B}u\, d\tau$$

Since any fundamental matrix can be written as a product of Φ and a constant matrix $\mathbf{C}$, we have

$$\begin{bmatrix} x_{11} & x_{12} \\ \dot{x}_{11} & \dot{x}_{12} \end{bmatrix} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

which is equivalent to

$$\Phi = \mathbf{X}\mathbf{C}^{-1} \qquad \text{or} \qquad \Phi^{-1} = \mathbf{C}\mathbf{X}^{-1}$$

where

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ \dot{x}_{11} & \dot{x}_{12} \end{bmatrix} \quad \text{and} \quad \mathbf{C} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

Then $\mathbf{x}(t)$ is obtained as

$$\mathbf{x}(t) = \mathbf{X}(t)\mathbf{C}^{-1} \int_{t_0}^{t} \mathbf{C}\mathbf{X}^{-1}(\tau)\mathbf{B}u \, d\tau$$

$$= \mathbf{X}(t) \int_{t_0}^{t} \mathbf{X}^{-1}(\tau)\mathbf{B}u \, d\tau$$

Notice that

$$\mathbf{X}^{-1} = \frac{1}{W(x_{11}, x_{12})} \begin{bmatrix} \dot{x}_{12} & -x_{12} \\ -\dot{x}_{11} & x_{11} \end{bmatrix}$$

Hence

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} x_{11}(t) & x_{12}(t) \\ \dot{x}_{11}(t) & \dot{x}_{12}(t) \end{bmatrix} \cdot \int_{t_0}^{t} \frac{1}{W(x_{11}, x_{12})} \begin{bmatrix} -x_{12}(\tau) \\ x_{11}(\tau) \end{bmatrix} u(\tau) \, d\tau$$

The solution $x(t)$ of Eq. (6–80) is thus given by

$$x(t) = x_1(t) = -x_{11}(t) \int_{t_0}^{t} \frac{x_{12}(\tau)u(\tau)}{W(x_{11}, x_{12})} \, d\tau + x_{12}(t) \int_{t_0}^{t} \frac{x_{11}(\tau)u(\tau)}{W(x_{11}, x_{12})} \, d\tau$$

Problem A-6-24 (*6–11*) Consider the system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \qquad \mathbf{x}(0) = \mathbf{0}$$

where $\mathbf{x}$ is an n vector, $\mathbf{u}$ is an r vector, $\mathbf{A}$ is an $n \times n$ constant matrix, and $\mathbf{B}$ is an $n \times r$ constant matrix. Suppose that the control vector $\mathbf{u}(t)$ is given by

$$\mathbf{u}(t) = \sum_{j=1}^{r} u_j(t) \, \mathcal{E}_j$$

where $\mathcal{E}_j$ is the unit vector which has a 1 in the jth place and 0's elsewhere, and $u_j(t)$ is given by

$$u_j(t) = \sum_{i=0}^{n-1} a_{ij} \overset{(i)}{\phi}(t)$$

$$(i = 0, 1, \ldots, n-1; j = 1, 2, \ldots, r)$$

where the a_{ij} are constants and $\overset{(i)}{\phi}(t)$ satisfies the condition that

$$\overset{(i)}{\phi}(0) = \overset{(i)}{\phi}(T) = 0 \qquad (i = 0, 1, 2, \ldots, n-2)$$

Show that

$$\mathbf{x}(T) = e^{\mathbf{A}T} \int_0^T e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}(\tau) \, d\tau$$

$$= e^{\mathbf{A}T} \left[\sum_{j=1}^{r} \sum_{i=0}^{n-1} a_{ij} \left(\int_0^T \phi(\tau) e^{-\mathbf{A}\tau} \, d\tau \right) \mathbf{A}^i \mathbf{B}_j \right]$$

where

$$\mathbf{B} = [\mathbf{B}_1 \mid \mathbf{B}_2 \mid \cdots \mid \mathbf{B}_r]$$

Solution By simple substitutions we obtain

$$\int_0^T e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}(\tau) \, d\tau = \sum_{j=1}^{r} \int_0^T e^{-\mathbf{A}\tau}\mathbf{B}u_j(\tau)\mathcal{E}_j \, d\tau$$

$$= \sum_{j=1}^{r} \sum_{i=0}^{n-1} a_{ij} \left(\int_0^T \overset{(i)}{\phi}(\tau) e^{-\mathbf{A}\tau} \, d\tau \right) \mathbf{B}_j \tag{6–82}$$

Notice that

$$\int_0^T \overset{(i)}{\phi}(\tau) e^{-\mathbf{A}\tau} \, d\tau = \left(\int_0^T \phi(\tau) e^{-\mathbf{A}\tau} \, d\tau \right) \mathbf{A}^i \tag{6–83}$$

$$(i = 0, 1, \ldots, n-1)$$

We can prove this by using induction on i and integration by parts. Clearly, Eq. (6–83) holds true for $i = 0$. If Eq. (6–83) is assumed true for i $(0 \leqslant i \leqslant n-2)$, then by integrating by parts, and noting that $\mathbf{A}$ and $e^{-\mathbf{A}}$ commute, we obtain

$$\int_0^T \overset{(i+1)}{\phi}(\tau) e^{-\mathbf{A}\tau} \, d\tau = e^{-\mathbf{A}\tau} \overset{(i)}{\phi}(\tau) \Big|_{\tau=0}^{\tau=T} + \int_0^T \overset{(i)}{\phi}(\tau)\mathbf{A}e^{-\mathbf{A}\tau} \, d\tau$$

$$= \left(\int_0^T \overset{(i)}{\phi}(\tau) e^{-\mathbf{A}\tau} \, d\tau \right) \mathbf{A} = \left(\int_0^T \phi(\tau) e^{-\mathbf{A}\tau} \, d\tau \right) \mathbf{A}^{i+1}$$

Hence, Eq. (6–82) becomes

$$\int_0^T e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}(\tau) \, d\tau = \sum_{j=1}^{r} \sum_{i=0}^{n-1} a_{ij} \left(\int_0^T \phi(\tau) e^{-\mathbf{A}\tau} \, d\tau \right) \mathbf{A}^i \mathbf{B}_j$$

and we obtain

$$\mathbf{x}(T) = e^{\mathbf{A}T} \left[\sum_{j=1}^{r} \sum_{i=0}^{n-1} a_{ij} \left(\int_0^T \phi(\tau) e^{-\mathbf{A}\tau} \, d\tau \right) \mathbf{A}^i \mathbf{B}_j \right]$$

Problem A-6-25 Consider the following system:

$$\mathbf{x}(k+1) = \mathbf{G}(k)\mathbf{x}(k) + \mathbf{H}(k)\mathbf{u}(k), \qquad \mathbf{x}(0) = \mathbf{x}_0$$

$$(k = 0, 1, 2, \ldots)$$

Show that the solution of this equation can be given by

$$\begin{aligned} \mathbf{x}(k) = {} & \mathbf{G}(k-1)\mathbf{G}(k-2) \cdots \mathbf{G}(0)\mathbf{x}_0 \\ & + \mathbf{G}(k-1)\mathbf{G}(k-2) \cdots \mathbf{G}(1)\mathbf{H}(0)\mathbf{u}(0) \\ & + \cdots \\ & + \mathbf{G}(k-1)\mathbf{H}(k-2)\mathbf{u}(k-2) \\ & + \mathbf{H}(k-1)\mathbf{u}(k-1) \end{aligned}$$

In particular, if $\mathbf{G}(k)$ and $\mathbf{H}(k)$ are constant matrices, then

$$\begin{aligned} \mathbf{x}(k) = {} & \mathbf{G}^k\mathbf{x}_0 + \mathbf{G}^{k-1}\mathbf{H}\mathbf{u}(0) + \mathbf{G}^{k-2}\mathbf{H}\mathbf{u}(1) \\ & + \cdots + \mathbf{G}^2\mathbf{H}\mathbf{u}(k-3) + \mathbf{G}\mathbf{H}\mathbf{u}(k-2) + \mathbf{H}\mathbf{u}(k-1) \end{aligned}$$

Solution The solution to the present difference equation is given by Eq. (6–66). By substituting $h = 0$ in Eq. (6–66) we obtain

$$\mathbf{x}(k) = \Phi(k, 0)\mathbf{x}_0 + \Phi(k, 0) \sum_{j=1}^{k} \Phi^{-1}(j, 0)\mathbf{H}(j-1)\mathbf{u}(j-1)$$

Since

$$\Phi(k, 0) = \mathbf{G}(k-1)\mathbf{G}(k-2) \cdots \mathbf{G}(0)$$

we have

$$\Phi(k, 0)\Phi^{-1}(j, 0) = \mathbf{G}(k-1)\mathbf{G}(k-2)\cdots\mathbf{G}(0)\mathbf{G}^{-1}(0)\mathbf{G}^{-1}(1)\cdots\mathbf{G}^{-1}(j-1)$$
$$= \mathbf{G}(k-1)\mathbf{G}(k-2)\cdots\mathbf{G}(j)$$

Hence,

$$\mathbf{x}(k) = \mathbf{G}(k-1)\mathbf{G}(k-2)\cdots\mathbf{G}(0)\mathbf{x}_0 + \mathbf{G}(k-1)\mathbf{G}(k-2)\cdots\mathbf{G}(1)\mathbf{H}(0)\mathbf{u}(0) + \cdots + \mathbf{G}(k-1)\mathbf{H}(k-2)\mathbf{u}(k-2) + \mathbf{H}(k-1)\mathbf{u}(k-1).$$

If G and **H** are constant matrices, then

$$\mathbf{x}(k) = \mathbf{G}^k\mathbf{x}_0 + \mathbf{G}^{k-1}\mathbf{H}\mathbf{u}(0) + \mathbf{G}^{k-2}\mathbf{H}\mathbf{u}(1) + \cdots + \mathbf{G}^2\mathbf{H}\mathbf{u}(k-3) + \mathbf{G}\mathbf{H}\mathbf{u}(k-2) + \mathbf{H}\mathbf{u}(k-1)$$

Problem A-6-26 Obtain the solution of the following difference equation:

$$x(k+2) + ax(k+1) + bx(k) = 0 \tag{6-84}$$

where a and b are constants and $k \geqslant 1$. The initial data are given as $x(0) = A$, $x(1) = B$.

Solution In this solution we shall present a classical method for solving difference equations. Let us define

$$Ex(k) = x(k+1)$$
$$E^2x(k) = x(k+2)$$

(E is a unit delay operator.) Then Eq. (6-84) becomes

$$(E^2 + aE + b)x(k) = 0$$

Let us put

$$x(k) = Cq^k \tag{6-85}$$

where C is an arbitrary constant and q is a number to be determined so that Eq. (6-85) is a solution of Eq. (6-84). Since

$$E^2x(k) = E^2Cq^k = q^2Cq^k = q^2x(k)$$

we have

$$q^2 + aq + b = 0 \tag{6-86}$$

or

$$q = m_1 \quad \text{or} \quad q = m_2$$

where m_1 and m_2 are roots of Eq. (6-86). Hence,

$$x(k) = C_1m_1^k + C_2m_2^k$$

The constants C_1 and C_2 are determined by

$$x(0) = A = C_1 + C_2$$
$$x(1) = B = C_1m_1 + C_2m_2$$

Hence,

$$C_1 = \frac{Am_2 - B}{m_2 - m_1}$$
$$C_2 = \frac{Am_1 - B}{m_1 - m_2}$$

Thus the solution $x(k)$ is obtained as

$$x(k) = \frac{Am_2 - B}{m_2 - m_1}m_1^k + \frac{Am_1 - B}{m_1 - m_2}m_2^k$$

Problem A-6-27 Consider the continuous-time system characterized by $\ddot{x} = u$. Discretize the system and derive a vector matrix difference equation for the system.

Solution By defining $x_1 = x$ and $x_2 = \dot{x}$, the given scalar differential equation becomes

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + \begin{bmatrix}0\\ 1\end{bmatrix}[u]$$

or

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u$$

The discrete-time equation is

$$\mathbf{x}(kT + \tau) = \mathbf{G}(\tau)\mathbf{x}(kT) + \mathbf{H}(\tau)u(kT), \qquad 0 < \tau \leqslant T$$

where T is the sampling period. The elements of $\mathbf{G}(\tau)$ correspond to the elements of

$$\mathscr{L}^{-1}[(s\mathbf{I} - \mathbf{A})^{-1}]$$

where

$$\mathbf{A} = \begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}$$

Since

$$s\mathbf{I} - \mathbf{A} = \begin{bmatrix}s & -1\\ 0 & s\end{bmatrix}$$

and

$$(s\mathbf{I} - \mathbf{A})^{-1} = \frac{1}{s^2}\begin{bmatrix}s & 1\\ 0 & s\end{bmatrix} = \begin{bmatrix}\frac{1}{s} & \frac{1}{s^2}\\ 0 & \frac{1}{s}\end{bmatrix}$$

we have

$$\mathbf{G}(\tau) = \mathscr{L}^{-1}[(s\mathbf{I} - \mathbf{A})^{-1}] = \begin{bmatrix}1 & \tau\\ 0 & 1\end{bmatrix}$$

The elements of $\mathbf{H}(\tau)$ are found from the expression

$$\mathbf{H}(\tau) = \int_0^\tau \mathbf{G}(\sigma)\mathbf{B}\,d\sigma$$

where

$$\mathbf{G}(\sigma)\mathbf{B} = \begin{bmatrix}1 & \sigma\\ 0 & 1\end{bmatrix}\begin{bmatrix}0\\ 1\end{bmatrix} = \begin{bmatrix}\sigma\\ 1\end{bmatrix}$$

Hence $\mathbf{H}(\tau)$ is obtained as

$$\mathbf{H}(\tau) = \begin{bmatrix} \frac{1}{2}\tau^2 \\ \tau \end{bmatrix}$$

Thus the discrete-time equation is

$$\begin{bmatrix} x_1(kT+\tau) \\ x_2(kT+\tau) \end{bmatrix} = \begin{bmatrix} 1 & \tau \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(kT) \\ x_2(kT) \end{bmatrix} + \begin{bmatrix} \frac{1}{2}\tau^2 \\ \tau \end{bmatrix} [u(kT)]$$

At the sampling instants, τ equals T. Therefore, we obtain

$$\begin{bmatrix} x_1((k+1)T) \\ x_2((k+1)T) \end{bmatrix} = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(kT) \\ x_2(kT) \end{bmatrix} + \begin{bmatrix} \frac{1}{2}T^2 \\ T \end{bmatrix} [u(kT)]$$

Problem A-6-28 The vector matrix differential equation

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{v} \tag{6-87}$$

can be discretized and the equation can be written as the vector matrix difference equation of the form

$$\mathbf{x}((k+1)T) = \mathbf{G}(T)\mathbf{x}(kT) + \mathbf{H}(T)\mathbf{u}(kT) + \mathbf{N}(T)\mathbf{v}(kT)$$

Determine $\mathbf{G}(T)$, $\mathbf{H}(T)$, and $\mathbf{N}(T)$. Assume that $\mathbf{A}$ and $\mathbf{B}$ are constant matrices, and $\mathbf{u}$ and $\mathbf{v}$ are constant between any two consecutive sampling instants.

Solution The formal solution of Eq. (6–87) is

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}_0 + e^{\mathbf{A}t}\int_0^t e^{-\mathbf{A}\tau}[\mathbf{B}\mathbf{u}(\tau) + \mathbf{v}(\tau)]\,d\tau$$

At $t = (k+1)T$, and $t = kT$ we have

$$\mathbf{x}((k+1)T) = e^{\mathbf{A}(k+1)T}\mathbf{x}_0 + e^{\mathbf{A}(k+1)T}\int_0^{(k+1)T} e^{-\mathbf{A}\tau} \cdot [\mathbf{B}\mathbf{u}(\tau) + \mathbf{v}(\tau)]d\tau \tag{6-88}$$

and

$$\mathbf{x}(kT) = e^{\mathbf{A}kT}\mathbf{x}_0 + e^{\mathbf{A}kT}\int_0^{kT} e^{-\mathbf{A}\tau}[\mathbf{B}\mathbf{u}(\tau) + \mathbf{v}(\tau)]\,d\tau \tag{6-89}$$

Multiplying Eq. (6–89) by $e^{\mathbf{A}T}$ and subtracting from Eq. (6–88), we obtain

$$\begin{aligned} \mathbf{x}((k+1)T) &= e^{\mathbf{A}T}\mathbf{x}(kT) + e^{\mathbf{A}(k+1)T}\int_{kT}^{(k+1)T} e^{-\mathbf{A}\tau}[\mathbf{B}\mathbf{u}(\tau) + \mathbf{v}(\tau)]\,d\tau \\ &= e^{\mathbf{A}T}\mathbf{x}(kT) + e^{\mathbf{A}T}\int_0^T e^{-\mathbf{A}t}\mathbf{B}\,dt\,\mathbf{u}(kT) + e^{\mathbf{A}T}\int_0^T e^{-\mathbf{A}t}\,dt\,\mathbf{v}(kT) \\ &= \mathbf{G}(T)\mathbf{x}(kT) + \mathbf{H}(T)\mathbf{u}(kT) + \mathbf{N}(T)\mathbf{v}(kT) \end{aligned}$$

Hence

$$\mathbf{G}(T) = e^{\mathbf{A}T}$$

$$\mathbf{H}(T) = e^{\mathbf{A}T}\int_0^T e^{-\mathbf{A}t}\mathbf{B}\,dt = \int_0^T e^{\mathbf{A}\tau}\mathbf{B}\,d\tau$$

$$\mathbf{N}(T) = e^{\mathbf{A}T}\int_0^T e^{-\mathbf{A}t}\,dt = \int_0^T e^{\mathbf{A}\tau}\,d\tau$$

PROBLEMS

Problem B-6-1 Obtain the fundamental matrix $\Phi(t)$ with $\Phi(0) = \mathbf{I}$ of the system

$$\overset{(n)}{x} = ax \qquad (n = 1, 2, 3, \ldots)$$

where a is a constant.

Problem B-6-2 Consider the following system:

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$$

where

$$\mathbf{A}(t) = \begin{bmatrix} a & e^{-t} \\ -e^{-t} & b \end{bmatrix}$$

a, b = constants

Obtain the fundamental matrix $\Phi(t, 0)$ with $\Phi(0, 0) = \mathbf{I}$.

Problem B-6-3 Assume that $\phi_1(t)$, $\phi_2(t)$, $\ldots$, $\phi_n(t)$ are any set of functions having at least up to the nth derivative with respect to t over the time interval $t_0 \leqslant t \leqslant t_1$. Assume also that the Wronskian does not vanish. Namely,

$$W(\phi_1, \phi_2, \ldots, \phi_n) = \begin{vmatrix} \phi_1 & \phi_2 & \cdots & \phi_n \\ \dot{\phi}_1 & \dot{\phi}_2 & \cdots & \dot{\phi}_n \\ \vdots & \vdots & & \vdots \\ \overset{(n-1)}{\phi_1} & \overset{(n-1)}{\phi_2} & \cdots & \overset{(n-1)}{\phi_n} \end{vmatrix} \neq 0$$

Show that

$$\frac{W(\phi_1, \phi_2, \ldots, \phi_n, x)}{W(\phi_1, \phi_2, \ldots, \phi_n)} = 0$$

is an nth-order linear scalar differential equation in x with the coefficient of the highest derivative term in x equal to unity.

Problem B-6-4 By use of Sylvester's interpolation formula, compute $e^{\mathbf{A}t}$ where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & -11 & -6 \end{bmatrix}$$

Problem B-6-5 Consider the following system:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Show that the fundamental matrix $\Phi(t)$ with $\Phi(t_0) = \mathbf{I}$ is given by

$$\Phi(t) = \begin{bmatrix} 2e^{-(t-t_0)} - e^{-2(t-t_0)} & e^{-(t-t_0)} - e^{-2(t-t_0)} \\ -2e^{-(t-t_0)} + 2e^{-2(t-t_0)} & -e^{-(t-t_0)} + 2e^{-2(t-t_0)} \end{bmatrix}$$

Problem B-6-6 Obtain the response $y(t)$ of the following system:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} -3 & -2 & -1 \\ 1 & 0 & -1 \\ 0 & 0 & -4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} [u]$$

$$y = x_1 + 2x_2 + 3x_3$$

where $u(t) = \delta(t) =$ unit impulse.

Problem B-6-7 Prove that the solution of the following set of n difference equations

$$x_i(k+1) = \sum_{j=1}^{n} g_{ij} x_j(k) + \sum_{m=1}^{r} h_{im} u_m(k)$$

$$(i = 1, 2, \ldots, n)$$

is given by

$$x_i(k) = \sum_{j=1}^{n} \phi_{ij}(k) x_j(0) + \sum_{s=0}^{k-1} \sum_{j=1}^{n} \sum_{m=1}^{r} \phi_{ij}(k-1-s) h_{jm} u_m(s)$$

$$= \sum_{j=1}^{n} \phi_{ij}(k) x_j(0) + \sum_{s=0}^{k-1} \sum_{j=1}^{n} \sum_{m=1}^{r} \phi_{ij}(s) h_{jm} u_m(k-1-s)$$

where the ϕ_{ij} are the solutions of the following equation:

$$\phi_{ij}(k+1) = \sum_{m=1}^{n} g_{im} \phi_{mj}(k) \qquad (i, j = 1, 2, \ldots, n)$$

with the initial conditions $\phi_{ij}(0) = \delta_{ij}$ where $\delta_{ij} = 0$ for $i \neq j$ and $\delta_{ij} = 1$ for $i = j$.

Note that this problem is the same as proving that the solution of

$$\mathbf{x}((k+1)T) = \mathbf{G}(T)\mathbf{x}(kT) + \mathbf{H}(T)\mathbf{u}(kT)$$

where $\mathbf{G}(T)$ and $\mathbf{H}(T)$ are constant matrices, is given by

$$\mathbf{x}(kT) = \mathbf{G}^k(T)\mathbf{x}(0) + \sum_{s=0}^{k-1} \mathbf{G}^s(T)\mathbf{H}(T)\mathbf{u}[(k-1-s)T]$$

$$(k = 0, 1, 2, 3, \ldots)$$

REFERENCES

6–1 Coddington, E. A., and N. Levinson, *Theory of Ordinary Differential Equations*, New York: McGraw-Hill Book Company, Inc., 1955.

6–2 Lefschetz, S., *Differential Equations: Geometric Theory*, New York: Interscience Publishers, Inc., 1957.

6–3 Seifert, W. W., and C. W. Steeg, Jr., *Control Systems Engineering*, New York: McGraw-Hill Book Company, Inc., 1960.

6–4 Enns, M., J. R. Greenwood III, J. E. Matheson, and F.T. Thompson, "Practical Aspects of State Space Methods: Part 1—System Formulation and Reduction; Part 2—System Analysis and Simulation," *JACC* Preprint (1964) 494–513.

6–5 Desoer, C. A.,"An Introduction to State Space Techniques in Linear Systems," *JACC* Preprint (1962), Paper No. 10–2.

6–6 Reid, W. T., "A Matrix Differential Equation of Riccati Type," *Amer. J. Math.*, **68** (1946), 237–46.

6–7 Butman, S., and R. Sivan (Sussman), "On Cancellations, Controllability and Observability," *IEEE Trans. Automatic Control* (1964), 317–18.

6–8 Gabasov, R., "On Optimal Processes in Coupled Digital Systems,"*Avtomatika i Telemekhanika*, **23**, No. 7 (July, 1962) 872–80 (Trans. "Automation and Remote Control," January, 1963, pp. 808–17.)

6–9 Ogata, K., Class Notes on Optimal Control Theory, Lecture, Yokohama National University, 1964–65.

6–10 Rogers, A. E., and T.W. Connolly, *Analog Computation in Engineering Design*, New York: McGraw-Hill Book Company, Inc., 1960, pp. 200–03.

6–11 Falb, P.L., and M. Athans, "A Direct Constructive Proof of the Criterion for Complete Controllability of Time-Invariant Linear Systems," *IEEE Trans. Automatic Control* (1964), 189–90.

7

Controllability and Observability

7-1 INTRODUCTION[1]

In solving basic problems of control of dynamic systems we must know the kind and amount of information necessary for achieving a desired control of a dynamic system. Consider the system shown in Fig. 7–1. The state and the output of the plant or dynamic system are described by the following equations:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$$
$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u} \qquad (7\text{–}1)$$

where

$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{y} = m$ vector (output vector)
$\mathbf{A} = n \times n$ matrix (transition matrix)
$\mathbf{B} = n \times r$ matrix (control matrix)
$\mathbf{C} = m \times n$ matrix (output matrix)
$\mathbf{D} = m \times r$ matrix (transmission matrix)

[1]For detailed exposition on the subject of controllability and observability see, for example, (*7–1*, *7–2*, *7–3*). Figures in parentheses refer to materials listed at the end of this chapter.

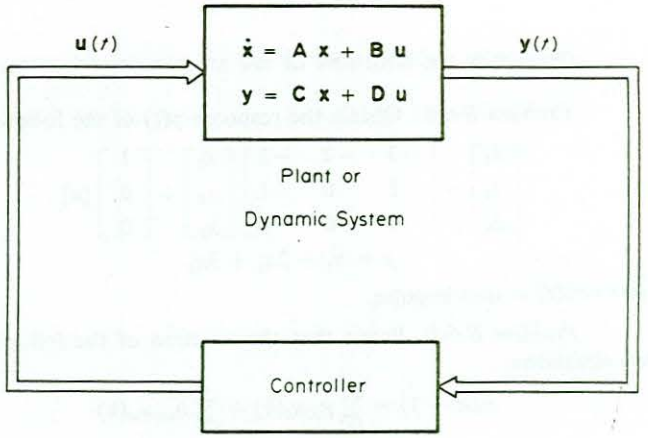


Figure 7-1 Control system.

A, **B**, **C**, and **D** are assumed to be constant matrices. The first equation of Eq. (7–1) gives the plant dynamics or system dynamics. The second equation of Eq. (7–1) specifies the output transformation. In Eq. (7–1) **A** characterizes the plant dynamics or system dynamics in the absence of **u**, **B** specifies how the plant or dynamic system is constrained with respect to **u**, **C** specifies how the output is constrained with respect to **x**, and **D** represents the direct transmission from **u** to **y**. [If every state **x** can be affected directly by **u**, the plant or dynamic system is not constrained with respect to **u**. If all the state variables $x_i(t)$ can be measured, the plant or dynamic system is not constrained with respect to **y**.] In the control system shown in Fig. 7–1, the controller observes the output $\mathbf{y}(t)$ and, based on this observation, determines the control vector $\mathbf{u}(t)$. Basic questions in problems of control of the plant or dynamic system may be stated as follows:

1. Is it possible to transform any initial state $\mathbf{x}(t_0)$ to any desired state in a finite length of time by applying appropriate control input?
2. Is it possible to identify the initial state $\mathbf{x}(t_0)$ by observing the output $\mathbf{y}(t)$ for a finite time interval?

Kalman introduced concepts of controllability and observability and gave answers to these basic questions (*7–1*). Our main concern in this chapter is giving a clear explanation of concepts of controllability and observability and answering the same questions.

Concepts of controllability and observability are of fundamental importance in modern control theory. In fact, controllability and observability may be interpreted as necessary and sometimes as sufficient conditions for the existence of a solution to most control problems.

In the following, we present definitions of controllability and observa-

bility. Note that hereafter in this chapter we shall call a plant or dynamic system which is described by Eq. (7–1) simply a *system*.

Controllability A system of Eq. (7–1) is said to be *completely state controllable* if for any t_0 it is possible to construct an unconstrained control vector $\mathbf{u}(t)$ which will transfer any given initial state $\mathbf{x}(t_0)$ to any final state $\mathbf{x}(t_1)$ in a finite time interval, $t_0 \leqslant t \leqslant t_1$. (We may conveniently take the final state to be the origin of the state space.)

Complete state controllability is neither necessary nor sufficient for the existence of a solution to the problem of controlling the output $\mathbf{y}(t)$ of the system. For this reason it is desirable to define separately complete output controllability. Output controllability allows the consideration of systems with time delays and with direct transmissions from control input to output (7–2).

A system of Eq. (7–1) is said to be *completely output controllable* if it is possible to construct an unconstrained control vector $\mathbf{u}(t)$ which will transfer any given initial output $\mathbf{y}(t_0)$ to any final output $\mathbf{y}(t_1)$ in a finite time interval $t_0 \leqslant t \leqslant t_1$.

Although complete controllability is generally a necessary condition for the existence of a solution to a control problem where t_1 is undefined, a stronger property of controllability is needed when the control process is confined to fixed finite intervals. Such stronger property of controllability is defined in terms of total controllability (7–2). A system is said to be *totally (state or output) controllable* if it is completely controllable on every interval $t_0 \leqslant t \leqslant t_1$.

Observability A system is said to be *completely observable* on $t_0 \leqslant t \leqslant t_1$ if, for every t_0 and some t_1, every state $\mathbf{x}(t_0)$ can be determined from the knowledge of $\mathbf{y}(t)$ on $t_0 \leqslant t \leqslant t_1$. Essentially, a system is completely observable if every transition of the system's state eventually affects the system's output.

From the foregoing definition complete observability means that once A and C of Eq. (7–1) are given and the response $\mathbf{y}(t)$ over a finite time interval is known, it is possible to determine the initial state of the system $\mathbf{x}(t_0)$. Consider a simple example. If the output transformation matrix C is given as

$$\mathbf{C} = [1 \quad 0 \quad 0 \quad \cdots \quad 0]$$

then in the absence of the control vector or $\mathbf{u}(t) = \mathbf{0}$ we have

$$y(t) = x_1(t)$$

In this case, the state variable $x_1(t)$ can be measured or observed but $x_2(t)$, $x_3(t)$, $\ldots$, $x_n(t)$ are neither measurable nor observable. If the initial values

of state variables $x_2(t_0)$, $x_3(t_0)$, $\ldots$, $x_n(t_0)$ can be determined on the basis of measured values of $x_1(t)$ over a finite time interval $t_0 \leqslant t \leqslant t_1$, then the system is completely observable. The concept of observability is useful in solving the problem of reconstructing unmeasurable state variables from measurable ones in the minimum possible length of time.

Similar to the case of total controllability we define *total observability* as follows: If the system is completely observable for every t_0 and every $t_1 > t_0$, the system is said to be totally observable.

Outline of the chapter This chapter presents a detailed exposition of concepts of, as well as theorems on, controllability and observability. Our discussions are limited to linear systems. Section 7–2 presents a mathematical background necessary for discussing controllability and observability of linear dynamic systems. Section 7–3 gives an elementary exposition of controllability of linear time-invariant systems. Section 7–4 presents detailed discussions of observability of linear time-invariant systems. Section 7–5 is concerned with controllability and observability of linear time-varying systems. Section 7–6 presents the relation between controllability and observability of dynamic systems and the relation between the linear dynamic system and its impulse response.

7-2 MATHEMATICAL BACKGROUND

In this section we present a mathematical background which we shall use in the rest of this chapter. First, we briefly explain the convex set; then we define the adjoint system. Use of adjoint systems often simplifies theoretical study of optimal control problems. We shall next discuss representation of elements in the n-dimensional vector space V relative to a given basis. Finally, we shall define Gram matrices and Gram determinants and present a few theorems on Gram matrices.

Convex set A set is said to be *convex* if all the points of the line segment joining every pair of points in the set are also in the set.

Let us define the set of all the points in the n-dimensional space which can be reached from the origin in the time interval $t_0 \leqslant t \leqslant t_1$ as $\Gamma(t_0, t_1)$. Then, $\Gamma(t_0, t_1)$ is a convex set. This can be seen as follows: Suppose that particular control vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ defined for $t_0 \leqslant t \leqslant t_1$ drive the output to points P_1 and P_2, respectively. Since any point in the n-dimensional vector space can be represented by an n-dimensional vector $\mathbf{x}$, let us represent points P_1 and P_2 by vectors $\mathbf{x}_1(t_1)$ and $\mathbf{x}_2(t_1)$, respectively. Consider any point P_3 on the line segment joining points P_1 and P_2. Let us represent point P_3 by a vector $\mathbf{x}_3(t_1)$. Then $\mathbf{x}_3(t_1)$ can be expressed as

$$\mathbf{x}_3(t_1) = c\mathbf{x}_1(t_1) + (1 - c)\mathbf{x}_2(t_1), \qquad 0 \leqslant c \leqslant 1$$

Clearly point P_3 can be reached by use of the control vector $\mathbf{u}_3$, where

$$\mathbf{u}_3 = c\mathbf{u}_1 + (1 - c)\mathbf{u}_2$$

Hence P_3 is in $\Gamma(t_0, t_1)$. Thus the set $\Gamma(t_0, t_1)$ is shown to be convex, since any point on the line segment joining every pair of points in $\Gamma(t_0, t_1)$ is also in $\Gamma(t_0, t_1)$.

If we wish to show that the convex set $\Gamma(t_0, t_1)$ coincides with the whole space V, we need show only that $\Gamma(t_0, t_1)$ is not bounded in any direction in V; in other words, the vector $\mathbf{x}(t_1)$ spans the n-dimensional space V.

Adjoint system Consider a system described by

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} \tag{7-2}$$

where $\mathbf{x}$ is an n vector and $\mathbf{A}(t)$ is an $n \times n$ matrix. We assume that the elements of $\mathbf{A}(t)$ are absolutely integrable as functions of t in the interval $t_0 \leqslant t \leqslant t_1$. Suppose that $\Phi(t, t_0)$ is the unique fundamental matrix satisfying

$$\frac{d}{dt}[\Phi(t, t_0)] = \mathbf{A}(t)\Phi(t, t_0), \qquad \Phi(t_0, t_0) = \mathbf{I} \tag{7-3}$$

Then differentiating both sides of

$$\Phi(t, t_0)\Phi^{-1}(t, t_0) = \mathbf{I}$$

with respect to t,

$$\begin{aligned}
&\frac{d}{dt}[\Phi(t, t_0)]\Phi^{-1}(t, t_0) + \Phi(t, t_0)\frac{d}{dt}[\Phi^{-1}(t, t_0)] \\
&\quad = \mathbf{A}(t)\Phi(t, t_0)\Phi^{-1}(t, t_0) + \Phi(t, t_0)\frac{d}{dt}[\Phi^{-1}(t, t_0)] \\
&\quad = \mathbf{A}(t) + \Phi(t, t_0)\frac{d}{dt}[\Phi^{-1}(t, t_0)] = \mathbf{0}
\end{aligned}$$

Hence, we obtain

$$\frac{d}{dt}[\Phi^{-1}(t, t_0)] = -\Phi^{-1}(t, t_0)\mathbf{A}(t), \qquad \Phi^{-1}(t_0, t_0) = \mathbf{I} \tag{7-4}$$

Let us define

$$\Phi(t_0, t) = \Phi^{-1}(t, t_0) = \Psi^*(t, t_0) \tag{7-5}$$

Then Eq. (7-4) becomes

$$\frac{d}{dt}[\Psi^*(t, t_0)] = -\Psi^*(t, t_0)\mathbf{A}(t) \tag{7-6}$$

Hence we see that $\Psi^*(t, t_0)$ is a fundamental matrix for the system

$$\dot{\mathbf{y}} = -\mathbf{y}\mathbf{A}(t)$$

where $\mathbf{y}$ is a row vector.

By taking the conjugate transpose of Eq. (7-6) we obtain

$$\frac{d}{dt}[\Psi(t, t_0)] = -\mathbf{A}^*(t)\Psi(t, t_0), \qquad \Psi(t_0, t_0) = \mathbf{I} \tag{7-7}$$

We find that $\Psi(t, t_0)$ is a fundamental matrix for the system

$$\dot{\mathbf{z}} = -\mathbf{A}^*(t)\mathbf{z} \tag{7-8}$$

Equation (7-8) is said to be *adjoint* to Eq. (7-2). The matrix differential equation (7-7) is said to be adjoint to Eq. (7-3). Conversely, Eq. (7-2) and Eq. (7-3) are said to be adjoint to Eq. (7-8) and Eq. (7-7), respectively.

If the matrix $\mathbf{A}(t)$ is skew Hermitian, or

$$\mathbf{A}(t) = -\mathbf{A}^*(t)$$

then the system

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$$

is said to be *self-adjoint*.

Example 7-1 Consider the following system:

$$\overset{(n)}{x} + a_1\overset{(n-1)}{x} + \cdots + a_n x = 0 \tag{7-9}$$

Let us obtain the adjoint system. Equation (7-9) can be written as

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x \\ \dot{x} \\ \vdots \\ \overset{(n-1)}{x} \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_1 \end{bmatrix}$$

The adjoint system is

$$\dot{\mathbf{y}} = -\mathbf{A}^*\mathbf{y}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad -\mathbf{A}^* = \begin{bmatrix} 0 & 0 & \cdots & 0 & \bar{a}_n \\ -1 & 0 & \cdots & 0 & \bar{a}_{n-1} \\ 0 & -1 & \cdots & 0 & \bar{a}_{n-2} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & \bar{a}_1 \end{bmatrix}$$

or

$$\begin{aligned}
\dot{y}_1 &= \bar{a}_n y_n \\
\dot{y}_2 &= -y_1 + \bar{a}_{n-1} y_n \\
&\cdots \quad \cdots \quad \cdots \\
\dot{y}_n &= -y_{n-1} + \bar{a}_1 y_n
\end{aligned} \tag{7-10}$$

By eliminating $y_1, y_2, \ldots, y_{n-1}$ from Eq. (7–10), we obtain

$$(-)^n y_n^{(n)} + (-)^{n-1}\bar{a}_1 y_n^{(n-1)} + \cdots + (-)\bar{a}_{n-1}\dot{y}_n + \bar{a}_n y_n = 0 \qquad (7\text{–}11)$$

Equation (7–11) is adjoint to Eq. (7–9).

Representation of elements in n-dimensional vector space relative to given basis Notice that from Eq. (7–5) we have

$$\Psi^*(t, t_0)\,\Phi(t, t_0) = \mathbf{I} \qquad \text{for all } t \text{ and } t_0$$

which is the same as

$$\begin{aligned}(\Psi_i(t, t_0),\ \Phi_j(t, t_0)) = \delta_{ij} &= 1 \qquad i = j \\ &= 0 \qquad i \neq j \\ (i, j &= 1, 2, \ldots, n)\end{aligned} \qquad (7\text{–}12)$$

where $\Psi_i(t, t_0)$ and $\Phi_j(t, t_0)$ are the ith column of $\Psi(t, t_0)$ and jth column of $\Phi(t, t_0)$, respectively. From Eq. (7–12) and since n columns of $\Phi(t, t_0)$ are linearly independent and n columns of $\Psi(t, t_0)$ are also linearly independent, a set of vectors $\Phi_i(t, t_0)$ $(i = 1, 2, \ldots, n)$ is a basis for the n-dimensional state space V and a set of vectors $\Psi_j(t, t_0)$ $(j = 1, 2, \ldots, n)$ is a basis for the dual state space V'. Therefore, referring to Section 2–3 we can express any vector $\mathbf{y}$ in V as

$$\mathbf{y}(t) = \sum_{i=1}^{n} (\Psi_i(t, t_0), \mathbf{y}(t))\,\Phi_i(t, t_0)$$

This can be seen easily. Since $\mathbf{y}(t)$ is expressible as

$$\mathbf{y}(t) = \sum_{i=1}^{n} a_i\,\Phi_i(t, t_0)$$

the inner product of $\Psi_i(t, t_0)$ and $\mathbf{y}(t)$ becomes

$$(\Psi_i(t, t_0),\ \mathbf{y}(t)) = \left(\Psi_i(t, t_0),\ \sum_{i=1}^{n} a_i\Phi_i(t, t_0)\right) = a_i$$

Hence,

$$\mathbf{y}(t) = \sum_{i=1}^{n} (\Psi_i(t, t_0), \mathbf{y}(t))\,\Phi_i(t, t_0) \qquad \text{for all } t \text{ and } t_0$$

We shall often use such representation of vectors in the rest of this chapter. For example, for a given basis $\xi_1, \xi_2, \ldots, \xi_n$ of the n-dimensional vector space V, we shall express any vector $\mathbf{y}$ of V uniquely as follows:

$$\mathbf{y} = \sum_{i=1}^{n} (\eta_i, \mathbf{y})\,\xi_i$$

where $\eta_1, \eta_2, \ldots, \eta_n$ are the dual basis of the basis $\xi_1, \xi_2, \ldots, \xi_n$.

Gram matrix and Gram determinant Suppose that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are elements of an m-dimensional vector space V. The $n \times n$ matrix

$$\mathbf{G} = ((\mathbf{x}_i, \mathbf{x}_j)) = \begin{bmatrix} (\mathbf{x}_1, \mathbf{x}_1) & (\mathbf{x}_1, \mathbf{x}_2) & \cdots & (\mathbf{x}_1, \mathbf{x}_n) \\ (\mathbf{x}_2, \mathbf{x}_1) & (\mathbf{x}_2, \mathbf{x}_2) & \cdots & (\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & & \vdots \\ (\mathbf{x}_n, \mathbf{x}_1) & (\mathbf{x}_n, \mathbf{x}_2) & \cdots & (\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}$$

is called the *Gram matrix* of the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Its determinant $|(\mathbf{x}_i, \mathbf{x}_j)|$ is called the *Gram determinant*.

Note that the Gram matrix can be represented as $\mathbf{A}^*\mathbf{A}$, since

$$\mathbf{G} = \begin{bmatrix} (\mathbf{x}_1, \mathbf{x}_1) & (\mathbf{x}_1, \mathbf{x}_2) & \cdots & (\mathbf{x}_1, \mathbf{x}_n) \\ (\mathbf{x}_2, \mathbf{x}_1) & (\mathbf{x}_2, \mathbf{x}_2) & \cdots & (\mathbf{x}_2, \mathbf{x}_n) \\ & \vdots & & \vdots \\ (\mathbf{x}_n, \mathbf{x}_1) & (\mathbf{x}_n, \mathbf{x}_2) & \cdots & (\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}$$

$$= \begin{bmatrix} \bar{x}_{11} & \bar{x}_{21} & \cdots & \bar{x}_{m1} \\ \bar{x}_{12} & \bar{x}_{22} & \cdots & \bar{x}_{m2} \\ \vdots & \vdots & & \vdots \\ \bar{x}_{1n} & \bar{x}_{2n} & \cdots & \bar{x}_{mn} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} = \mathbf{A}^*\mathbf{A}$$

where

$$\mathbf{A} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

The $n \times n$ matrix $\mathbf{G}$ is also called the Gram matrix of the $m \times n$ matrix $\mathbf{A}$. (Clearly, if $\mathbf{A}$ is a real matrix, then $\mathbf{A}^* = \mathbf{A}'$ and the Gram matrix can be written as $\mathbf{A}'\mathbf{A}$.)

We shall next prove that the Gram matrix is always positive definite or positive semidefinite.

Theorem 7-1 The $n \times n$ Gram matrix $\mathbf{A}^*\mathbf{A}$ of an $m \times n$ matrix $\mathbf{A}$ is positive definite if the rank of $\mathbf{A}$ is n, and it is positive semidefinite if the rank of $\mathbf{A}$ is less than n.

Proof Consider the Hermitian form

$$\begin{aligned}\mathbf{x}^*(\mathbf{A}^*\mathbf{A})\mathbf{x} &= (\mathbf{A}\mathbf{x})^*(\mathbf{A}\mathbf{x}) = \mathbf{y}^*\mathbf{y} \\ &= \bar{y}_1 y_1 + \bar{y}_2 y_2 + \cdots + \bar{y}_m y_m \\ &= |y_1|^2 + |y_2|^2 + \cdots + |y_m|^2 \geqslant 0\end{aligned}$$

Notice that $\mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x}$ becomes equal to zero if and only if $\mathbf{y} = \mathbf{A}\mathbf{x} = \mathbf{0}$. If the rank of $\mathbf{A}$ is n, then the only vector $\mathbf{x}$ that satisfies $\mathbf{A}\mathbf{x} = \mathbf{0}$ is $\mathbf{x} = \mathbf{0}$. Hence the Gram matrix $\mathbf{A}^*\mathbf{A}$ is positive definite if the rank of $\mathbf{A}$ is n.

Nonzero vectors $\mathbf{x}$ that satisfy $\mathbf{A}\mathbf{x} = \mathbf{0}$ exist if and only if the rank of $\mathbf{A}$ is less than n. Hence the Gram matrix $\mathbf{A}^*\mathbf{A}$ is positive semidefinite if the rank of $\mathbf{A}$ is less than n. This completes the proof.

Notice that any positive definite matrices and positive semidefinite matrices can be represented as Gram matrices. (Refer to Prob. A-7-7.)

Theorem 7-2 A necessary and sufficient condition that real or complex m vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be linearly independent is that the Gram matrix $((\mathbf{x}_i, \mathbf{x}_j))$ be nonsingular, or the following Gram determinant be nonzero.

$$|(\mathbf{x}_i, \mathbf{x}_j)| = \begin{vmatrix} (\mathbf{x}_1, \mathbf{x}_1) & (\mathbf{x}_1, \mathbf{x}_2) & \cdots & (\mathbf{x}_1, \mathbf{x}_n) \\ (\mathbf{x}_2, \mathbf{x}_1) & (\mathbf{x}_2, \mathbf{x}_2) & \cdots & (\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & & \vdots \\ (\mathbf{x}_n, \mathbf{x}_1) & (\mathbf{x}_n, \mathbf{x}_2) & \cdots & (\mathbf{x}_n, \mathbf{x}_n) \end{vmatrix}$$

Proof To prove this, suppose that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are linearly dependent. Then, there exist constants $c_1, c_2, \ldots, c_n$ not all zero such that

$$c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_n\mathbf{x}_n = \mathbf{0}$$

Hence,

$$\sum_{j=1}^{n} c_j(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i, \sum_{j=1}^{n} c_j\mathbf{x}_j) = 0$$

This means that column vectors of the $n \times n$ matrix $((\mathbf{x}_i, \mathbf{x}_j))$ become linearly dependent. Hence,

$$|(\mathbf{x}_i, \mathbf{x}_j)| = 0$$

Conversely, if

$$|(\mathbf{x}_i, \mathbf{x}_j)| = 0$$

then there exist constants $c_1, c_2, \ldots, c_n$ not all zero such that

$$\sum_{j=1}^{n} c_j(\mathbf{x}_i, \mathbf{x}_j) = 0 \qquad (i = 1, 2, \ldots, n)$$

Hence,

$$(\sum_{i=1}^{n} c_i\mathbf{x}_i, \sum_{j=1}^{n} c_j\mathbf{x}_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} \bar{c}_i c_j(\mathbf{x}_i, \mathbf{x}_j) = 0$$

it follows that

$$\sum_{i=1}^{n} c_i\,\mathbf{x}_i = \mathbf{0}$$

which means that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are linearly dependent. Hence, we have proved that a necessary and sufficient condition that vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are linearly independent is

$$|(\mathbf{x}_i, \mathbf{x}_j)| \neq 0$$

which implies that the Gram matrix is nonsingular. This completes the proof. (Note that the Gram determinant is nonzero implies that it is positive. See Prob. A-7-6.)

Linear independence of column vectors which are continuous functions of time Consider next the following $m \times n$ matrix $\mathbf{H}(t, t_0)$:

$$\mathbf{H}(t, t_0) = \begin{bmatrix} h_{11}(t, t_0) & h_{12}(t, t_0) & \cdots & h_{1n}(t, t_0) \\ h_{21}(t, t_0) & h_{22}(t, t_0) & \cdots & h_{2n}(t, t_0) \\ \vdots & \vdots & & \vdots \\ h_{m1}(t, t_0) & h_{m2}(t, t_0) & \cdots & h_{mn}(t, t_0) \end{bmatrix}$$

$$= [\mathbf{H}_1(t, t_0) \mid \mathbf{H}_2(t, t_0) \mid \cdots \mid \mathbf{H}_n(t, t_0)]$$

where the functions $h_{ij}(t, t_0)$ $(i = 1, 2, \ldots, m; j = 1, 2, \ldots, n)$ are continuous functions of t for the interval $t_0 \leqslant t \leqslant t_1$. The vectors $\mathbf{H}_1(t, t_0)$, $\mathbf{H}_2(t, t_0)$, $\ldots$, $\mathbf{H}_n(t, t_0)$ are called *linearly independent* if and only if

$$c_1\mathbf{H}_1(t, t_0) + c_2\mathbf{H}_2(t, t_0) + \cdots + c_n\mathbf{H}_n(t, t_0) = \mathbf{0} \qquad (7\text{–}13)$$

where $c_1, c_2, \ldots, c_n$ are constants, implies $c_1 = c_2 = \cdots = c_n = 0$. The vectors $\mathbf{H}_1(t, t_0)$, $\mathbf{H}_2(t, t_0)$, $\ldots$, $\mathbf{H}_n(t, t_0)$ are called *linearly dependent* if there exist constants $c_1, c_2, \ldots, c_n$ not all zero such that Eq. (7–13) holds true.

For time-varying vectors $\mathbf{H}_i(t, t_0)$ and $\mathbf{H}_j(t, t_0)$, we define the inner product thus:

$$(\mathbf{H}_i, \mathbf{H}_j) = \int_{t_0}^{t_1} \mathbf{H}_i^*(t, t_0)\,\mathbf{H}_j(t, t_0)\,dt$$

In Section 2–3 it was stated that any definition of inner product must satisfy the four axioms. Clearly, the present definition satisfies the four axioms, namely

1. $(\mathbf{H}_i, \mathbf{H}_j) = \overline{(\mathbf{H}_j, \mathbf{H}_i)}$
2. $(c\mathbf{H}_i, \mathbf{H}_j) = \bar{c}(\mathbf{H}_i, \mathbf{H}_j) = (\mathbf{H}_i, \bar{c}\mathbf{H}_j)$
   where c is a complex number
3. $(\mathbf{H}_i + \mathbf{H}_j, \mathbf{H}_h + \mathbf{H}_k) = (\mathbf{H}_i, \mathbf{H}_h) + (\mathbf{H}_j, \mathbf{H}_h) + (\mathbf{H}_i, \mathbf{H}_k) + (\mathbf{H}_j, \mathbf{H}_k)$
4. $(\mathbf{H}_i, \mathbf{H}_i) > 0$ for $\mathbf{H}_i \neq \mathbf{0}$

We shall define the norm $\|\mathbf{H}_j\|$ of $\mathbf{H}_j = \mathbf{H}_j(t, t_0)$ by $(\mathbf{H}_j, \mathbf{H}_j)^{1/2}$.

The Gram matrix of $\mathbf{H}(t, t_0)$ is given by

$$\begin{bmatrix} (\mathbf{H}_1, \mathbf{H}_1) & (\mathbf{H}_1, \mathbf{H}_2) & \cdots & (\mathbf{H}_1, \mathbf{H}_n) \\ (\mathbf{H}_2, \mathbf{H}_1) & (\mathbf{H}_2, \mathbf{H}_2) & \cdots & (\mathbf{H}_2, \mathbf{H}_n) \\ \vdots & \vdots & & \vdots \\ (\mathbf{H}_n, \mathbf{H}_1) & (\mathbf{H}_n, \mathbf{H}_2) & \cdots & (\mathbf{H}_n, \mathbf{H}_n) \end{bmatrix}$$

$$= \begin{bmatrix} \int_{t_0}^{t_1} \mathbf{H}_1^* \mathbf{H}_1 dt & \int_{t_0}^{t_1} \mathbf{H}_1^* \mathbf{H}_2 dt & \cdots & \int_{t_0}^{t_1} \mathbf{H}_1^* \mathbf{H}_n dt \\ \int_{t_0}^{t_1} \mathbf{H}_2^* \mathbf{H}_1 dt & \int_{t_0}^{t_1} \mathbf{H}_2^* \mathbf{H}_2 dt & \cdots & \int_{t_0}^{t_1} \mathbf{H}_2^* \mathbf{H}_n dt \\ \vdots & \vdots & & \vdots \\ \int_{t_0}^{t_1} \mathbf{H}_n^* \mathbf{H}_1 dt & \int_{t_0}^{t_1} \mathbf{H}_n^* \mathbf{H}_2 dt & \cdots & \int_{t_0}^{t_1} \mathbf{H}_n^* \mathbf{H}_n dt \end{bmatrix}$$

$$= \int_{t_0}^{t_1} \begin{bmatrix} \mathbf{H}_1^* \mathbf{H}_1 & \mathbf{H}_1^* \mathbf{H}_2 & \cdots & \mathbf{H}_1^* \mathbf{H}_n \\ \mathbf{H}_2^* \mathbf{H}_1 & \mathbf{H}_2^* \mathbf{H}_2 & \cdots & \mathbf{H}_2^* \mathbf{H}_n \\ \vdots & \vdots & & \vdots \\ \mathbf{H}_n^* \mathbf{H}_1 & \mathbf{H}_n^* \mathbf{H}_2 & \cdots & \mathbf{H}_n^* \mathbf{H}_n \end{bmatrix} dt$$

$$= \int_{t_0}^{t_1} \mathbf{H}^*(t, t_0)\, \mathbf{H}(t, t_0)\, dt$$

Theorem 7-3 A set of vectors $\mathbf{H}_1(t, t_0), \mathbf{H}_2(t, t_0), \ldots, \mathbf{H}_n(t, t_0)$, where

$$\mathbf{H}_j(t, t_0) = \begin{bmatrix} h_{1j}(t, t_0) \\ h_{2j}(t, t_0) \\ \vdots \\ h_{mj}(t, t_0) \end{bmatrix} \qquad (j = 1, 2, \ldots, n)$$

and the elements $h_{ij}(t, t_0)$ $(i = 1, 2, \ldots, m; j = 1, 2, \ldots, n)$ are continuous functions of t over the time interval $t_0 \leqslant t \leqslant t_1$, are linearly independent over the same time interval if and only if the Gram matrix

$$\int_{t_0}^{t_1} \mathbf{H}^*(t, t_0)\, \mathbf{H}(t, t_0)\, dt$$

of the $m \times n$ matrix

$$\mathbf{H}(t, t_0) = [\mathbf{H}_1(t, t_0) \mid \mathbf{H}_2(t, t_0) \mid \cdots \mid \mathbf{H}_n(t, t_0)]$$

is nonsingular, or the Gram determinant

$$\left| (\mathbf{H}_i, \mathbf{H}_j) \right| = \left| \int_{t_0}^{t_1} \mathbf{H}_i^*(t, t_0)\, \mathbf{H}_j(t, t_0)\, dt \right|$$

The proof of this theorem is similar to that of Theorem 7–2 and is, therefore, omitted. We shall use this theorem in discussing controllability and observability of linear time-varying systems.

From Theorem 7–3 we can immediately obtain the following result: A necessary and sufficient condition for $x_1(t), x_2(t), \ldots, x_n(t)$, a set of n real or complex valued continuous functions of t defined over the interval $t_0 \leqslant t \leqslant t_1$, to be linearly independent is that the Gram determinant

$$\left| (x_i, x_j) \right| = \left| \int_{t_0}^{t_1} x_i^*(t)\, x_j(t) dt \right|$$

be nonzero.

7-3 CONTROLLABILITY OF LINEAR TIME-INVARIANT SYSTEMS

In this section we shall discuss controllability of linear time-invariant systems. We shall treat systems of the form

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$$
$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}$$

where
$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{y} = m$ vector (output vector) $(m \leqslant n)$
$\mathbf{A} = n \times n$ matrix
$\mathbf{B} = n \times r$ matrix
$\mathbf{C} = m \times n$ matrix
$\mathbf{D} = m \times r$ matrix

or their equivalent discrete-time systems. The controllability of linear time-varying systems is discussed in Section 7–5.

In the rest of this section, we first consider complete state controllability of a simple discrete-time system and determine an optimal control law that will bring every state of the system to the origin in a minimal number of sampling periods. Then, we shall consider complete state controllability and complete output controllability of continuous-time systems.

Consider a simple discrete-time linear time-invariant system with a single control signal. For time-invariant systems, the initial time t_0 can be chosen zero. Assume that the equation describing the system is given by

$$\mathbf{x}((k+1)T) = \mathbf{G}[\mathbf{x}(kT) + \mathbf{F}u(kT)] \tag{7-14}$$

where
$\mathbf{x} = n$ vector (state vector)
$u =$ scalar (control signal)
$\mathbf{G} = n \times n$ nonsingular matrix
$\mathbf{F} = n \times 1$ matrix
$T =$ sampling period

The desired state of the system is assumed to be identically equal to zero for all $t \geqslant kT$. The system is state controllable if there exists a control signal $u(t)$ defined over a finite time interval $0 \leqslant t \leqslant kT$ such that starting from any initial state $\mathbf{x}(0)$ the state $\mathbf{x}(t)$ is zero for $t \geqslant kT$ or $\mathbf{x}(t \geqslant kT) = \mathbf{0}$. If every state is controllable, the system is completely state controllable. If this system is completely state controllable, it is possible to bring the arbitrary initial state $\mathbf{x}(0)$ to the origin in a finite time interval $0 \leqslant t \leqslant kT$.

We shall first make a preliminary investigation for obtaining a necessary and sufficient condition for complete state controllability. Referring to Eq. (7–14), we have

$$\mathbf{x}(T) = \mathbf{G}[\mathbf{x}(0) + \mathbf{F}u(0)]$$

By setting $\mathbf{x}(T) = \mathbf{0}$ we can obtain a set of initial states $\mathbf{x}(0)$ from which we can bring the system state to the origin in one sampling period. Namely,

$$\mathbf{0} = \mathbf{G}[\mathbf{x}(0) + \mathbf{F}u(0)] \qquad \text{or} \qquad \mathbf{x}(0) = -\mathbf{F}u(0)$$

If the magnitude of the control signal $u(0)$ is not bounded, then the initial state $\mathbf{x}(0)$ lying on the vector $\mathbf{F}$ can be brought to the origin in one sampling period. A set of initial states $\mathbf{x}(0)$ from which we can bring the system state to the origin in two sampling periods can be obtained as follows: From Eq. (7–14)

$$\mathbf{x}(2T) = \mathbf{G}[\mathbf{x}(T) + \mathbf{F}u(T)]$$

By setting $\mathbf{x}(2T) = \mathbf{0}$, we obtain

$$\begin{aligned} \mathbf{0} &= \mathbf{G}[\mathbf{x}(T) + \mathbf{F}u(T)] \\ &= \mathbf{G}[\mathbf{G}\mathbf{x}(0) + \mathbf{G}\mathbf{F}u(0) + \mathbf{F}u(T)] \end{aligned}$$

Since $\mathbf{G}$ is nonsingular,

$$\mathbf{x}(0) = -\mathbf{F}u(0) - \mathbf{G}^{-1}\mathbf{F}u(T)$$

Similarly, a set of initial states $\mathbf{x}(0)$, from which we can transfer the system state to the origin in n sampling periods, can be obtained as follows:

$$\begin{aligned} \mathbf{x}(0) = &- \mathbf{F}u(0) - \mathbf{G}^{-1}\mathbf{F}u(T) - \mathbf{G}^{-2}\mathbf{F}u(2T) \\ &- \cdots - \mathbf{G}^{-n+1}\mathbf{F}u((n-1)T) \end{aligned} \tag{7–15}$$

Let us denote $\mathbf{G}^{-i+1}\mathbf{F}$ by $\boldsymbol{\xi}_i$ $(i = 1, 2, \ldots, n)$. It can be shown by induction that the set $\Gamma(q)$ of all initial states which can be transferred to the origin in at most q steps by the application of an appropriate control signal is given by

$$\Gamma(q) = \{\mathbf{x}; \mathbf{x} = \sum_{i=1}^{q} \beta_i \boldsymbol{\xi}_i\}$$

where the β_i are arbitrary numbers. Note that if $\boldsymbol{\xi}_{k+1}$ is linearly dependent on $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_k$, then so is $\boldsymbol{\xi}_m$ for every $m > k$. Hence if $\Gamma(k) = \Gamma(k+1)$, then $\Gamma(k) = \Gamma(m)$ for all $m \geqslant k$.

A necessary and sufficient condition for complete state controllability in the discrete-time case can be stated in the form of a theorem as follows:

Theorem 7-4 *(7–1)* A discrete-time system described by Eq. (7–14), rewritten as

$$\mathbf{x}((k+1)T) = \mathbf{G}[\mathbf{x}(kT) + \mathbf{F}u(kT)]$$

is completely state controllable if and only if the composite $n \times n$ matrix $\mathbf{P}$ where

$$\mathbf{P} = [\mathbf{F} \mid \mathbf{G}\mathbf{F} \mid \cdots \mid \mathbf{G}^{n-1}\mathbf{F}]$$

is of rank n. (Notice that since $\mathbf{G}$ is nonsingular this condition is equivalent to stating that the composite $n \times n$ matrix $\mathbf{Q}$ where

$$\mathbf{Q} = [\mathbf{F} \mid \mathbf{G}^{-1}\mathbf{F} \mid \cdots \mid \mathbf{G}^{-n+1}\mathbf{F}]$$

is of rank n.)

Proof Notice first that the matrix $\mathbf{Q}$ has rank n means that the vectors $\mathbf{F}, \mathbf{G}^{-1}\mathbf{F}, \ldots, \mathbf{G}^{-n+1}\mathbf{F}$ are linearly independent. If the set of vectors $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_n$ where $\boldsymbol{\xi}_i = \mathbf{G}^{-i+1}\mathbf{F}$ is linearly independent, it is a basis in the state space of the system. Any initial state $\mathbf{x}(0)$ can be written as a linear combination of the vectors $\boldsymbol{\xi}_i$. [See Eq. (7–15).] Hence, every state can be transferred to the origin in at most n steps or n sampling periods and so the system is completely state controllable.

Conversely, suppose that only k (where $k < n$) of the vectors $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_n$ are linearly independent but the system is completely state controllable. Then the first k of these vectors are linearly independent and, therefore, $\Gamma(k) = \Gamma(m)$ for all $m \geqslant k$. Hence $\Gamma(k)$ is the set of all controllable states. But, the dimension of $\Gamma(k)$ is k which is less than n. This contradicts the assumption that the system is completely state controllable. Therefore, the n vectors $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_n$ must be linearly independent and the rank of $\mathbf{Q}$ must be n. This completes the proof.

We shall next consider the problem of finding a control signal u which will transfer the state $\mathbf{x}$ starting from an arbitrary initial state $\mathbf{x}(0)$ to the origin in some finite time T. Specifically, we shall solve the problem of finding an optimal control law that will bring every state of a discrete-time system described by Eq. (7–14) to the origin in a minimal number of sampling periods. We shall present this problem and solution in the form of a theorem.

Theorem 7-5 *(7–1)* Consider a discrete-time completely state controllable system described by

$$\mathbf{x}((k+1)T) = \mathbf{G}[\mathbf{x}(kT) + \mathbf{F}u(kT)]$$

where

$\mathbf{x} = n$ vector (state vector)
$u =$ scalar (control signal)
$\mathbf{G} = n \times n$ nonsingular matrix
$\mathbf{F} = n \times 1$ matrix
$T =$ sampling period

Define

$$\xi_i = \mathbf{G}^{-i+1}\mathbf{F} \qquad (i = 1, 2, \ldots, n)$$

Every state of this system can be transferred to the origin of the state space in a minimal number (n or less) of sampling periods if and only if the control law is given by

$$u(kT) = -(\eta_1, \mathbf{x}(kT)) \tag{7-16}$$

where η_1 is the first vector in the dual basis $\eta_1, \eta_2, \ldots, \eta_n$ of the basis $\xi_1, \xi_2, \ldots, \xi_n$ of the n-dimensional state space.

Proof Let us denote the set of all initial states by $\Gamma(m)$. Then at any time kT, any vector in $\Gamma(m)$ (where $1 \leqslant m \leqslant n$) can be represented in the following form:

$$\mathbf{x}(kT) = \sum_{i=1}^{m} \beta_i \xi_i = \sum_{i=1}^{m} (\eta_i, \mathbf{x}(kT))\xi_i$$

Suppose that $\beta_m \neq 0$, that is, that $\mathbf{x}(kT)$ is in $\Gamma(m)$ but not in $\Gamma(m-1)$. Using the control law given by Eq. (7–16) and the relationship $\mathbf{G}\xi_{i+1} = \xi_i$, we obtain

$$\mathbf{x}((k+1)T) = \mathbf{G}\sum_{i=2}^{m} (\eta_i, \mathbf{x}(kT))\xi_i$$

$$= \sum_{i=1}^{m-1} (\eta_{i+1}, \mathbf{x}(kT))\xi_i = \sum_{i=1}^{m-1} (\eta_i, \mathbf{x}((k+1)T))\xi_i$$

We thus see that the control law given by Eq. (7–16) is the only control law assuring that $\mathbf{x}((k+1)T)$ will be in $\Gamma(m-1)$. Hence we have proved the theorem.

Notice that the control law given by Eq. (7–16) satisfies the principle of feedback, since the control signal is computed from the measurement of the system state $\mathbf{x}(kT)$. Note that the control law, Eq. (7–16), may not be physically realizable, since constraints may prevent the instantaneous measurement of all state variables.

We shall next extend Theorem 7–4 to the case where $\mathbf{u}$ is an r vector.

Theorem 7-6 A discrete-time system described by

$$\mathbf{x}((k+1)T) = \mathbf{G}[\mathbf{x}(kT) + \mathbf{F}\mathbf{u}(kT)] \tag{7-17}$$

where
$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{G} = n \times n$ nonsingular matrix
$\mathbf{F} = n \times r$ matrix
$T =$ sampling period

is completely state controllable if and only if the composite $n \times nr$ matrix $\mathbf{P}$ where

$$\mathbf{P} = [\mathbf{F} \mid \mathbf{GF} \mid \cdots \mid \mathbf{G}^{n-1}\mathbf{F}]$$

is of rank n. (The matrix $\mathbf{P}$ has nr columns.) The proof of this theorem is omitted since it is quite similar to those of Theorems 7–4 and 7–7.

We shall next consider the case where the system is continuous-time and linear time-invariant. The equation of the system is

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \tag{7-18}$$

where
$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{A} = n \times n$ matrix
$\mathbf{B} = n \times r$ matrix

Complete state controllability of the system can be stated by the following theorem:

Theorem 7-7 (*7–1, 7–4*) A continuous-time system described by Eq. (7–18) is completely state controllable if and only if the composite $n \times nr$ matrix $\mathbf{P}$ where

$$\mathbf{P} = [\mathbf{B} \mid \mathbf{AB} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{B}]$$

is of rank n. (The matrix $\mathbf{P}$ has nr columns.)

Proof The proof of this theorem is similar to that of the discrete-time case.

Suppose that the system is state controllable and the initial state $\mathbf{x}(0)$ can be transferred to the origin of the state space in a finite time interval $0 \leqslant t \leqslant T$. Since the solution of Eq. (7–18) is

$$\mathbf{x}(t) = e^{\mathbf{A}t}[\mathbf{x}(0) + \int_0^t e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}(\tau)d\tau]$$

at $t = T$ we have

$$\mathbf{0} = e^{\mathbf{A}T}\mathbf{x}(0) + \int_0^T e^{\mathbf{A}(T-\tau)}\mathbf{B}\mathbf{u}(\tau)d\tau$$

Noting that $e^{\mathbf{A}T}$ is nonsingular,

$$\mathbf{x}(0) = -\int_0^T e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}(\tau)d\tau$$

Referring to Section 6–4 we can write $e^{-\mathbf{A}\tau}$ in the following form:

$$e^{-\mathbf{A}\tau} = \sum_{i=0}^{n-1} \alpha_i \mathbf{A}^i$$

where p is the degree of the minimal polynomial of $\mathbf{A}$. Since the α_i are scalar valued functions of τ, let us explicitly indicate this dependency on τ and write

$$e^{-\mathbf{A}\tau} = \sum_{i=0}^{p-1} \alpha_i(\tau)\, \mathbf{A}^i$$

Then we obtain

$$\mathbf{x}(0) = -\sum_{i=0}^{p-1} \int_0^T \alpha_i(\tau)\mathbf{A}^i\mathbf{B}\mathbf{u}(\tau)d\tau \tag{7-19}$$

Since $\mathbf{u}(\tau)$ may be written as

$$\mathbf{u}(\tau) = \sum_{j=1}^{r} u_j(\tau)\boldsymbol{\epsilon}_j$$

where $\boldsymbol{\epsilon}_j$ is the unit vector which has a 1 in the jth place and 0's elsewhere, Eq. (7–19) can be written as follows:

$$\mathbf{x}(0) = -\sum_{i=0}^{p-1}\sum_{j=1}^{r}\left(\int_0^T \alpha_i(\tau)u_j(\tau)d\tau\right)\mathbf{A}^i\mathbf{B}_j$$

where $\mathbf{B}_j$ denotes the jth column of $\mathbf{B}$. Let us define

$$\int_0^T \alpha_i(\tau)u_j(\tau)d\tau = \beta_{ij} = \text{scalar}$$

then we have

$$\mathbf{x}(0) = -\sum_{i=0}^{p-1}\sum_{j=1}^{r}\beta_{ij}\mathbf{A}^i\mathbf{B}_j \tag{7-20}$$

From Eq. (7–20) we see that any initial state $\mathbf{x}(0)$ which can be transferred to the origin of the state space is a linear combination of $\mathbf{B}_j, \mathbf{AB}_j, \ldots, \mathbf{A}^{p-1}\mathbf{B}_j$ $(j = 1, 2, \ldots, r)$. Hence if the rank of the matrix $\mathbf{Q}$, where

$$\begin{aligned}\mathbf{Q} &= [\mathbf{B}_1 \mid \mathbf{B}_2 \mid \cdots \mid \mathbf{A}^{p-1}\mathbf{B}_{r-1} \mid \mathbf{A}^{p-1}\mathbf{B}_r] \\ &= [\mathbf{B} \mid \mathbf{AB} \mid \cdots \mid \mathbf{A}^{p-1}\mathbf{B}]\end{aligned}$$

is n (meaning that there is a set of n linearly independent vectors which consititute a basis for the n-dimensional state space), then the system is completely state controllable. If $p = n$, then $\mathbf{Q} = \mathbf{P}$. Therefore we can state that if the rank of $\mathbf{P}$ is n, the system is completely state controllable. Notice that if $p < n$, the $\mathbf{A}^h\mathbf{B}_j$, where $h = p, p+1, \ldots, n-1$, are linearly dependent on $\mathbf{B}_j, \ldots, \mathbf{A}^{p-1}\mathbf{B}_j$. Hence, in the case of $p < n$, we have the same conclusion; that is, if the rank of $\mathbf{P}$ is n, then the system is completely state controllable.

Conversely, suppose that the system is completely state controllable but the rank of $\mathbf{P}$ is k, where $k < n$. Then, the set $\Gamma(k)$ of all initial states which can be transferred to the origin is of k-dimensional space. Hence the dimension of $\Gamma(k)$ is less than n. This contradicts the assumption that the system is completely state controllable. Therefore the rank of $\mathbf{P}$ must be n. This completes the proof.

Notice that if the system equation is given by

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{B}u$$

where $\mathbf{x}$ is an n vector, u is a scalar, and

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_1 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

then it is easily verified that

$$\mathbf{B}, \mathbf{AB}, \ldots, \mathbf{A}^{n-1}\mathbf{B}$$

are linearly independent. Hence all systems of this type are completely state controllable.

Example 7-2 Consider the following system:

$$\begin{bmatrix}\dot{x}_1 \\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}-3 & 1 \\ -2 & 1.5\end{bmatrix}\begin{bmatrix}x_1 \\ x_2\end{bmatrix} + \begin{bmatrix}0 \\ 1\end{bmatrix}[u]$$

Let us examine state controllability of the system.

In this system

$$\mathbf{A} = \begin{bmatrix}-3 & 1 \\ -2 & 1.5\end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix}0 \\ 1\end{bmatrix}$$

Hence,

$$\mathbf{AB} = \begin{bmatrix}-3 & 1 \\ -2 & 1.5\end{bmatrix}\begin{bmatrix}0 \\ 1\end{bmatrix} = \begin{bmatrix}1 \\ 1.5\end{bmatrix}$$

We see that vectors $\mathbf{B}$ and $\mathbf{AB}$ are linearly independent and the rank of the matrix $[\mathbf{B} \mid \mathbf{AB}]$ is two. The system is, therefore, completely state controllable.

Example 7-3 Discuss state controllability of the following system:

$$\begin{bmatrix}\dot{x}_1 \\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}-3 & 1 \\ -2 & 1.5\end{bmatrix}\begin{bmatrix}x_1 \\ x_2\end{bmatrix} + \begin{bmatrix}1 \\ 4\end{bmatrix}[u] \tag{7-21}$$

In this system

$$\mathbf{A} = \begin{bmatrix}-3 & 1 \\ -2 & 1.5\end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix}1 \\ 4\end{bmatrix}$$

Since

$$\mathbf{AB} = \begin{bmatrix}-3 & 1 \\ -2 & 1.5\end{bmatrix}\begin{bmatrix}1 \\ 4\end{bmatrix} = \begin{bmatrix}1 \\ 4\end{bmatrix}$$

we see that vectors $\mathbf{B}$ and $\mathbf{AB}$ are not independent and the rank of the matrix $[\mathbf{B} \mid \mathbf{AB}]$ is one. Therefore, the system is not completely state controllable. In fact, elimination of x_2 from Eq. (7–21), or the following two simultaneous differential equations

$$\dot{x}_1 = -3x_1 + x_2 + u$$
$$\dot{x}_2 = -2x_1 + 1.5x_2 + 4u$$

yields

$$\ddot{x}_1 + 1.5\dot{x}_1 - 2.5x_1 = \dot{u} + 2.5u$$

or, in the form of the transfer function,

$$\frac{X_1(s)}{U(s)} = \frac{s + 2.5}{(s + 2.5)(s - 1)}$$

It is clear that cancellation of the factor $(s + 2.5)$ occurs in the numerator and denominator of the transfer function. (Thus one degree of freedom gets lost.) Because of this cancellation this system is not completely state controllable.

Consider the nth-order system whose transfer function is

$$G(s) = \frac{p(s)}{q(s)}$$

where $p(s)$ and $q(s)$ are of orders m and n $(m < n)$, respectively. If $p(s)$ and $q(s)$ have a factor or factors in common and cancellation occurs, then the order of the system becomes $n - k$ $(0 < k \leqslant m)$ which is lower than n. Such a system is not completely state controllable. This can be seen as follows: Suppose that the system is completely state controllable. Then the set $\Gamma(n - k)$ of all initial states which can be transferred to the origin is of $(n - k)$-dimensional space. The dimension of $\Gamma(n - k)$ is less than n. This contradicts the fact that the system is of order n.

A necessary and sufficient condition for complete state controllability is that no cancellation occurs in the transfer function or transfer matrix. (Stated differently, the transfer function or transfer matrix has some cancellations if the system is not completely state controllable.) In the following we shall consider in detail complete state controllability in the s domain.

Consider the linear time-invariant system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u, \qquad \mathbf{x}(0) = \mathbf{0} \tag{7-22}$$

where
$\mathbf{x} = n$ vector (state vector)
$u =$ scalar (control signal)
$\mathbf{A} = n \times n$ matrix
$\mathbf{B} = n \times 1$ matrix

Laplace transforming Eq. (7–22) and solving for $\mathbf{X}(s)$ we obtain

$$\mathbf{X}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}U(s) \tag{7-23}$$

where

$$(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} = \frac{1}{|s\mathbf{I} - \mathbf{A}|}\begin{bmatrix} p_1(s) \\ p_2(s) \\ \vdots \\ p_n(s) \end{bmatrix}$$

and the $p_i(s)$ $(i = 1, 2, \ldots, n)$ are polynomials in s. We shall next define what we mean by *cancellations* in the transfer matrix $(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$. The matrix $(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$ is said to have no cancellation if and only if the polynomials $p_1(s), p_2(s), \ldots, p_n(s)$, and $|s\mathbf{I} - \mathbf{A}|$ have no common factor. If $(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$ has a cancellation, then the system cannot be controlled in the direction of the canceled mode. The following theorem gives a necessary and sufficient condition for cancellations in the transfer matrix.

Theorem 7-8 (7–5) Consider an n vector $(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$,

where
$\mathbf{A} = n \times n$ matrix
$\mathbf{B} = n \times 1$ matrix
$\mathbf{I} = n \times n$ identity matrix
$s =$ complex variable such that $(s\mathbf{I} - \mathbf{A})^{-1}$ exists

$(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$ has a cancellation if and only if the rank of

$$\mathbf{P} = [\mathbf{B} \mid \mathbf{AB} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{B}]$$

is less than n.

Proof Let us define

$$\varphi \equiv (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$$

By the use of Eq. (6–38) we can express φ as follows:

$$\varphi = \frac{\sum_{j=0}^{n-1} s^j \sum_{i=1+j}^{n} \alpha_i \mathbf{A}^{i-j-1}\mathbf{B}}{\sum_{i=0}^{n} \alpha_i s^i} \qquad (\alpha_n = 1)$$

Define

$$\mathbf{v}_j = \sum_{i=1+j}^{n} \alpha_i \mathbf{A}^{i-j-1}\mathbf{B}$$

where $\mathbf{v}_j$ is an n vector. Then

$$\varphi = \frac{\sum_{j=0}^{n-1} s^j \mathbf{v}_j}{\sum_{i=0}^{n} \alpha_i s^i} \tag{7-24}$$

Suppose that φ has a cancellation. Then the numerator of the right-hand side of Eq. (7–24) must have the following form:

$$\sum_{j=0}^{n-1} s^j \mathbf{v}_j = (s - s_k) \sum_{j=0}^{n-2} s^j \mathbf{w}_j \tag{7-25}$$

where s_k is an eigenvalue of $\mathbf{A}$. By equating the coefficients of s^i $(i = 0, 1, 2, \ldots, n - 1)$ of both sides of Eq. (7–25) we obtain

$$\begin{aligned}
\mathbf{v}_0 &= -s_k\mathbf{w}_0 \\
\mathbf{v}_1 &= \mathbf{w}_0 - s_k\mathbf{w}_1 \\
\mathbf{v}_2 &= \mathbf{w}_1 - s_k\mathbf{w}_2 \\
&\cdots \quad \cdots \\
\mathbf{v}_{n-1} &= \mathbf{w}_{n-2}
\end{aligned}$$

Then

$$\begin{aligned}
&\mathbf{v}_0 + s_k\mathbf{v}_1 + s_k^2\mathbf{v}_2 + \cdots + s_k^{n-1}\mathbf{v}_{n-1} \\
&= (-s_k\mathbf{w}_0) + (s_k\mathbf{w}_0 - s_k^2\mathbf{w}_1) + (s_k^2\mathbf{w}_1 - s_k^3\mathbf{w}_2) + \cdots \\
&\quad + (s_k^{n-2}\mathbf{w}_{n-3} - s_k^{n-1}\mathbf{w}_{n-2}) + s_k^{n-1}\mathbf{w}_{n-2} \\
&= \mathbf{0}
\end{aligned} \tag{7-26}$$

Equation (7–26) implies that

$$\sum_{i=1}^{n} \alpha_i\mathbf{A}^{i-1}\mathbf{B} + s_k\sum_{i=2}^{n} \alpha_i\mathbf{A}^{i-2}\mathbf{B} + \cdots + s_k^{n-1}\alpha_n\mathbf{B} = \mathbf{0}$$

which can be rewritten as

$$c_0\mathbf{B} + c_1\mathbf{AB} + c_2\mathbf{A}^2\mathbf{B} + \cdots + c_{n-1}\mathbf{A}^{n-1}\mathbf{B} = \mathbf{0} \tag{7-27}$$

where,

$$c_i = \sum_{j=0}^{n-i-1} \alpha_{i+j+1}s_k^j$$

Since the coefficient of $\mathbf{A}^{n-1}\mathbf{B}$ is $c_{n-1} = \alpha_n = 1$, Eq. (7–27) implies that the vectors $\mathbf{B}, \mathbf{AB}, \ldots, \mathbf{A}^{n-1}\mathbf{B}$ are linearly dependent. Hence the rank of $\mathbf{P}$ is less than n. Thus we have proved that if φ has a cancellation, then the rank of $\mathbf{P}$ is less than n.

We shall next prove that if the rank of $\mathbf{P}$ is less than n, then φ has a cancellation. If the rank of $\mathbf{P}$ is less than n, then there exist constants $\gamma_0, \gamma_1, \ldots, \gamma_{n-1}$, not all zero, such that

$$\gamma_0\mathbf{B} + \gamma_1\mathbf{AB} + \gamma_2\mathbf{A}^2\mathbf{B} + \cdots + \gamma_{n-1}\mathbf{A}^{n-1}\mathbf{B} = \mathbf{0}$$

Since

$$\mathbf{B} = (s\mathbf{I} - \mathbf{A})\varphi$$

we have

$$\sum_{i=0}^{n-1} \gamma_i\mathbf{A}^i(s\mathbf{I} - \mathbf{A})\varphi = (s\mathbf{I} - \mathbf{A})\sum_{i=0}^{n-1} \gamma_i\mathbf{A}^i\varphi = \mathbf{0} \tag{7-28}$$

For s such that $|s\mathbf{I} - \mathbf{A}| \neq 0$, Eq. (7–28) implies that

$$\sum_{i=0}^{n-1} \gamma_i\mathbf{A}^i\varphi = \mathbf{0}$$

Using the identities

$$\begin{aligned}
\varphi &= \varphi \\
\mathbf{A}\varphi &= s\varphi - \mathbf{B} \\
\mathbf{A}^2\varphi &= s^2\varphi - \mathbf{AB} - s\mathbf{B} \\
&\cdots \quad \cdots \quad \cdots \\
\mathbf{A}^{n-1}\varphi &= s^{n-1}\varphi - \mathbf{A}^{n-2}\mathbf{B} - s\mathbf{A}^{n-3}\mathbf{B} - \cdots - s^{n-2}\mathbf{B}
\end{aligned}$$

we obtain

$$\begin{aligned}
\sum_{i=0}^{n-1} \gamma_i\mathbf{A}^i\varphi &= \sum_{i=0}^{n-1} \gamma_i s^i\varphi - \sum_{i=1}^{n-1} \gamma_i\mathbf{A}^{i-1}\mathbf{B} - s\sum_{i=2}^{n-1} \gamma_i\mathbf{A}^{i-2}\mathbf{B} \\
&\quad - \cdots - s^{n-3}\sum_{i=n-2}^{n-1} \gamma_i\mathbf{A}^{i-n+2}\mathbf{B} - s^{n-2}\gamma_{n-1}\mathbf{B} \\
&= \mathbf{0}
\end{aligned}$$

Hence,

$$\sum_{i=0}^{n-1} \gamma_i s^i\varphi = \sum_{j=0}^{n-2} s^j \sum_{i=j+1}^{n-1} \gamma_i\mathbf{A}^{i-j-1}\mathbf{B}$$

or

$$\varphi = \frac{\sum_{j=0}^{n-2} s^j \sum_{i=j+1}^{n-1} \gamma_i\mathbf{A}^{i-j-1}\mathbf{B}}{\sum_{i=0}^{n-1} \gamma_i s^i} \tag{7-29}$$

The denominator of Eq. (7–29) indicates that a cancellation occurred. This completes the proof.

We have shown that the condition that the rank of the matrix

$$[\mathbf{B} \mid \mathbf{AB} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{B}]$$

is n is equivalent to the condition that no cancellation occurs in the transfer matrix $(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$.

Theorem 7-9 A necessary and sufficient condition that the system defined by

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{B}u$$

where

$\mathbf{x} = n$ vector (state vector)
$u =$ scalar (control signal)
$\mathbf{A} = n \times n$ matrix
$\mathbf{B} = n \times 1$ matrix

is completely state controllable is that $(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$ has no cancellation.

Proof The proof of this theorem follows from Theorems 7-7 and 7–8. We leave details to the reader.

Example 7-4 Consider the system defined by the following transfer function:

$$\frac{X(s)}{U(s)} = \frac{K(s+a)(s+b)}{(s+a)(s+b)(s+c)(s+d)}$$

The transfer function has cancellations and, therefore, the system is not completely state controllable. The simplified state space representation of this system when a, b, c, and d are distinct is

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \\ \dot{z}_4 \end{bmatrix} = \begin{bmatrix} -a & 0 & 0 & 0 \\ 0 & -b & 0 & 0 \\ 0 & 0 & -c & 0 \\ 0 & 0 & 0 & -d \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} [u]$$

Observe that because of cancellations in the transfer function the motions of z_1 and z_2 do not depend on u. (Therefore z_1 and z_2 cannot be controlled.)

From this example we see that the distinct eigenvalue system is completely state controllable if and only if all the components of the control matrix in the simplified state space representation (meaning that the transition matrix is in the diagonal form) are different from zero.

Example 7-5 Consider the system defined by

$$\frac{X(s)}{U(s)} = \frac{K}{(s+a)^2(s+c)(s+d)}$$

where a, c, and d are distinct. The simplified state space representation of this system is

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \\ \dot{z}_4 \end{bmatrix} = \begin{bmatrix} -a & 1 & 0 & 0 \\ 0 & -a & 0 & 0 \\ 0 & 0 & -c & 0 \\ 0 & 0 & 0 & -d \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} [u]$$

This system is completely state controllable.

Consider next the system defined by

$$\frac{X(s)}{U(s)} = \frac{K(s+a)^2(s+d)}{(s+a)^2(s+c)(s+d)}$$

where a, c, and d are distinct. The simplified state space representation of this system is

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \\ \dot{z}_4 \end{bmatrix} = \begin{bmatrix} -a & 1 & 0 & 0 \\ 0 & -a & 0 & 0 \\ 0 & 0 & -c & 0 \\ 0 & 0 & 0 & -d \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} [u]$$

This system is not completely state controllable, because of cancellations in the transfer function, or stated differently because the motions of z_1, z_2, and z_4 are uncoupled with u.

Note that if in the control matrix (the coefficient matrix of u) any of the components corresponding to the distinct eigenvalues or corresponding to the last row of the Jordan block (for the multiple eigenvalues) is zero, then the system is not completely state controllable.

Complete state controllability of the linear time-invariant system defined by either

$$\mathbf{x}((k+1)T) = \mathbf{G}[\mathbf{x}(kT) + \mathbf{F}u(kT)]$$

or

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u$$

can be stated in various ways. We shall give below four definitions which are equivalent. (*7–1*)

1 The system is completely state controllable if and only if a set of vectors $\mathbf{F}, \mathbf{GF}, \ldots, \mathbf{G}^{n-1}\mathbf{F}$ or a set of vectors $\mathbf{B}, \mathbf{AB}, \ldots, \mathbf{A}^{n-1}\mathbf{B}$ spans the n-dimensional state space.
2 The system is completely state controllable if and only if there are no cancellations in the transfer function or transfer matrix.
3 The system is completely state controllable if and only if there exists a sampled data controller which has "deadbeat" response.
4 The system is completely state controllable if and only if (a) in the Jordan canonical form of $\mathbf{G}$ or $\mathbf{A}$ no two blocks are associated with the same eigenvalues, and (b) in the control matrix (the coefficient matrix of u) all the components corresponding to the distinct eigenvalues and those corresponding to the last row of each Jordan block (for multiple eigenvalues) are different from zero.

To examine condition (a) note that if two or more blocks in the Jordan canonical form of $\mathbf{G}$ or $\mathbf{A}$ are associated with the same eigenvalues, then the minimal polynomial is of lower degree than n. Hence the rank of

$$[\mathbf{F} \mid \mathbf{GF} \mid \cdots \mid \mathbf{G}^{n-1}\mathbf{F}]$$

or

$$[\mathbf{B} \mid \mathbf{AB} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{B}]$$

is lower than n. Thus in such a case the system is not completely state controllable. Condition (b) follows directly from discussions given in Examples 7–4 and 7–5.

Consider next a stable or asymptotically stable system. If the norm of $\mathbf{u}(t)$ is bounded due to saturation, then it may take longer to bring the state $\mathbf{x}(t)$ starting from the initial state $\mathbf{x}(0)$ to the origin. If, however, the system is neither stable nor asymptotically stable, then the region of controllability with bounded $\mathbf{u}(t)$ will not coincide with the n-dimensional state space V, but is a subspace of V. (Refer to Prob. A-7-12.) In general, we can say that if the norm of $\mathbf{u}(t)$ is bounded, not all initial states can necessarily be taken to the origin. The set $\Gamma_1(q)$ of all initial states that can be transferred to the origin in the presence of the constraints on the control vector $\mathbf{u}(t)$ is a subset of the set $\Gamma(q)$ of all initial states that can be transferred to the origin in the absence of the constraints on $\mathbf{u}(t)$.

In the case of a discrete-time system with a quantized control vector $\mathbf{u}(t)$,

only a finite number of specific initial states can be transferred to the origin. Often, however, we allow a small region, called δ region, around the origin and consider the system is acceptable if we can transfer the state to this δ region in a finite time interval $0 \leqslant t \leqslant kT$. Such a system may be said to be *δ state controllable.*

We shall next consider complete output controllability of systems in the form of theorems.

Theorem 7-10 The system described by

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{Ax} + \mathbf{Bu} \\ \mathbf{y} &= \mathbf{Cx} \end{aligned} \tag{7-30}$$

where

$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{y} = m$ vector (output vector) $(m \leqslant n)$
$\mathbf{A} = n \times n$ matrix
$\mathbf{B} = n \times r$ matrix
$\mathbf{C} = m \times n$ matrix

is completely output controllable if and only if the composite $m \times nr$ matrix $\mathbf{P}$ where

$$\mathbf{P} = [\mathbf{CB} \mid \mathbf{CAB} \mid \mathbf{CA}^2\mathbf{B} \mid \cdots \mid \mathbf{CA}^{n-1}\mathbf{B}]$$

is of rank m. (Notice that complete state controllability is neither necessary nor sufficient for complete output controllability.)

Proof Suppose that the system is output controllable and the output $\mathbf{y}(t)$ starting from any $\mathbf{y}(0)$, the initial output, can be transferred to the origin of the output space in a finite time interval $0 \leqslant t \leqslant T$. Namely,

$$\mathbf{y}(T) = \mathbf{Cx}(T) = \mathbf{0} \tag{7-31}$$

Since the solution of the first part of Eq. (7–30) is

$$\mathbf{x}(t) = e^{\mathbf{A}t}\left[\mathbf{x}(0) + \int_0^t e^{-\mathbf{A}\tau}\mathbf{Bu}(\tau)d\tau\right]$$

at $t = T$, we have

$$\mathbf{x}(T) = e^{\mathbf{A}T}\left[\mathbf{x}(0) + \int_0^T e^{-\mathbf{A}\tau}\mathbf{Bu}(\tau)d\tau\right] \tag{7-32}$$

Substituting Eq. (7–32) into Eq. (7–31) we obtain

$$\begin{aligned} \mathbf{y}(T) &= \mathbf{Cx}(T) \\ &= \mathbf{C}e^{\mathbf{A}T}\left[\mathbf{x}(0) + \int_0^T e^{-\mathbf{A}\tau}\mathbf{Bu}(\tau)d\tau\right] = \mathbf{0} \end{aligned} \tag{7-33}$$

On the other hand $\mathbf{y}(0) = \mathbf{Cx}(0)$. Notice that the complete output controllability means that the vector $\mathbf{Cx}(0)$ spans the m-dimensional output space. Since $e^{\mathbf{A}T}$ is nonsingular, if $\mathbf{Cx}(0)$ spans the m-dimensional output space, so does $\mathbf{C}e^{\mathbf{A}T}\mathbf{x}(0)$, and vice versa. From Eq. (7–33) we obtain

$$\begin{aligned} \mathbf{C}e^{\mathbf{A}T}\mathbf{x}(0) &= -\mathbf{C}e^{\mathbf{A}T}\int_0^T e^{-\mathbf{A}\tau}\mathbf{Bu}(\tau)d\tau \\ &= -\mathbf{C}\int_0^T e^{\mathbf{A}\tau}\mathbf{Bu}(T-\tau)d\tau \end{aligned}$$

Similar to the case of Theorem 7–7 we can express $\int_0^T e^{\mathbf{A}\tau}\mathbf{Bu}(T-\tau)\,d\tau$ as a sum of $\mathbf{A}^i\mathbf{B}_j$ as follows:

$$\begin{aligned} &\int_0^T e^{\mathbf{A}\tau}\mathbf{Bu}(T-\tau)d\tau \\ &\quad = \sum_{i=0}^{p-1}\sum_{j=1}^{r}\gamma_{ij}\mathbf{A}^i\mathbf{B}_j \end{aligned}$$

where

$$\gamma_{ij} = \int_0^T \alpha_i(\tau)u_j(T-\tau)d\tau = \text{scalar}$$

and $\alpha_i(\tau)$ satisfies

$$e^{\mathbf{A}\tau} = \sum_{i=0}^{p-1}\alpha_i(\tau)\,\mathbf{A}^i \qquad (p\text{: degree of the minimal polynomial of } \mathbf{A})$$

and $\mathbf{B}_j$ is the jth column of $\mathbf{B}$. Therefore, we can write $\mathbf{C}e^{\mathbf{A}T}\mathbf{x}(0)$ as

$$\mathbf{C}e^{\mathbf{A}T}\mathbf{x}(0) = -\sum_{i=0}^{p-1}\sum_{j=1}^{r}\gamma_{ij}\mathbf{CA}^i\mathbf{B}_j$$

From this last equation we see that $\mathbf{C}e^{\mathbf{A}T}\mathbf{x}(0)$ is a linear combination of $\mathbf{CA}^i\mathbf{B}_j$ $(i = 0, 1, 2, \ldots, p-1; j = 1, 2, \ldots, r)$. Note that if the rank of $\mathbf{Q}$ where

$$\mathbf{Q} = [\mathbf{CB} \mid \mathbf{CAB} \mid \mathbf{CA}^2\mathbf{B} \mid \cdots \mid \mathbf{CA}^{p-1}\mathbf{B}]$$

is m, then so is the rank of $\mathbf{P}$, and vice versa. [This is obvious if $p = n$. If $p < n$ then the $\mathbf{CA}^h\mathbf{B}_j$ (where $p \leqslant h \leqslant n-1$) are linearly dependent on $\mathbf{CB}_j$, $\mathbf{CAB}_j$, $\ldots$, $\mathbf{CA}^{p-1}\mathbf{B}_j$. Hence the rank of $\mathbf{P}$ is equal to that of $\mathbf{Q}$.] If the rank of $\mathbf{P}$ is m, then $\mathbf{C}e^{\mathbf{A}T}\mathbf{x}(0)$ spans the m-dimensional output space. This means that if the rank of $\mathbf{P}$ is m, then $\mathbf{Cx}(0)$ also spans the m-dimensional output space and the system is completely output controllable.

Conversely, suppose that the system is completely output controllable but the rank of $\mathbf{P}$ is k, where $k < m$. Then the set of all initial outputs which can be transferred to the origin is of k-dimensional space. Hence the dimension of this set is less than m. This contradicts the assumption that the system is completely output controllable. This completes the proof.

From Theorem 7–10 it can be immediately proved that, in the system of Eq. (7–30), complete state controllability on $0 \leqslant t \leqslant T$ implies complete output controllability on $0 \leqslant t \leqslant T$ if and only if m rows of $\mathbf{C}$ are linearly independent.

Finally, we shall present a theorem on complete output controllability of a system involving a direct transmission from control input to output.

Theorem 7-11 (*7–2*) A continuous-time system described by

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}$$
$$\mathbf{y} = \mathbf{Cx} + \mathbf{Du}$$

where
$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{y} = m$ vector (output vector)
$\mathbf{A} = n \times n$ matrix
$\mathbf{B} = n \times r$ matrix
$\mathbf{C} = m \times n$ matrix
$\mathbf{D} = m \times r$ matrix

is completely output controllable if and only if the composite $m \times (n + 1)r$ matrix $\mathbf{P}$ where

$$\mathbf{P} = [\mathbf{CB} \mid \mathbf{CAB} \mid \mathbf{CA}^2\mathbf{B} \mid \cdots \mid \mathbf{CA}^{n-1}\mathbf{B} \mid \mathbf{D}]$$

is of rank m.

Proof The proof is similar to that of Theorem 7–10. We leave details to the reader.

Note that $\mathbf{D}$ indicates the direct transmission from $\mathbf{u}$ to $\mathbf{y}$. The existence of such a direct transmission from $\mathbf{u}$ to $\mathbf{y}$ does not affect state controllability, but it always aids output controllability.

7-4 OBSERVABILITY OF LINEAR TIME-INVARIANT SYSTEMS

The notion of observability of a dynamic system is associated with the processing of data obtained from observations on the system. The basic question is: given a mathematical model of a free dynamic system, $\dot{\mathbf{x}} = \mathbf{Ax}$, and the output transformation, $\mathbf{y} = \mathbf{Cx}$, what is a necessary and sufficient condition for us to determine the system state at any time t by observing the output over a finite time interval $t_0 \leqslant t \leqslant t_0 + T$? (If the linear time-invariant system is observable in a time interval $0 \leqslant t \leqslant T$, then it is observable in any time interval $0 \leqslant t \leqslant T_1$, where $T_1 > 0$.) The answer to this question is given as theorems in this section.

Notice that in discussing observability of systems we need consider only free dynamic systems. The reason for this is explained as follows: Consider a linear continuous-time system described by

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu} \tag{7–34}$$
$$\mathbf{y} = \mathbf{Cx}$$

where
$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{y} = m$ vector (output vector) $(m \leqslant n)$
$\mathbf{A} = n \times n$ matrix
$\mathbf{B} = n \times r$ matrix
$\mathbf{C} = m \times n$ matrix

The system is said to be completely observable if and only if the initial state $\mathbf{x}(0)$ can be determined from measurements of the output $\mathbf{y}(t)$ over a finite time interval $0 \leqslant t \leqslant T$. Since $\mathbf{x}(t)$ can be determined from Eq. (7–34) as

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}(0) + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{Bu}(\tau)d\tau$$

we obtain

$$\mathbf{y}(t) = \mathbf{C}\left[e^{\mathbf{A}t}\mathbf{x}(0) + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{Bu}(\tau)d\tau\right]$$

Since matrices $\mathbf{A}$ and $\mathbf{B}$ are known and $\mathbf{u}(t)$ is also a known quantity, the integral term on the right-hand side of this equation is a known quantity. Therefore, this quantity may be subtracted from the observed value of $\mathbf{y}(t)$. Hence we may write

$$\tilde{\mathbf{y}}(t) = \mathbf{y}(t) - \mathbf{C}\int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{Bu}(\tau)d\tau = \mathbf{C}e^{\mathbf{A}t}\mathbf{x}(0)$$

Therefore, for investigating a necessary and sufficient condition for observability it suffices to consider the following system which has no input $\mathbf{u}(t)$.

$$\dot{\mathbf{x}} = \mathbf{Ax} \tag{7–35}$$
$$\mathbf{y} = \mathbf{Cx}$$

where $\mathbf{x}$, $\mathbf{y}$, $\mathbf{A}$, and $\mathbf{C}$ are defined in Eq. (7–34).

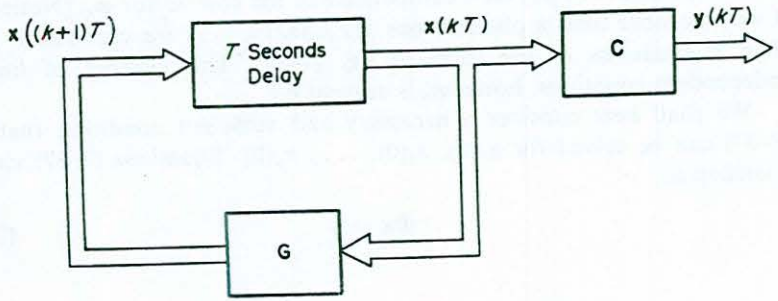


Figure 7-2 Linear discrete-time system.

Consider the system shown in Fig. 7–2. To simplify our analysis let us assume the system has a single-output. Equations of the system are then given by

$$\begin{aligned} \mathbf{x}((k+1)T) &= \mathbf{G}\mathbf{x}(kT) \\ y(kT) &= \mathbf{C}\mathbf{x}(kT) \end{aligned} \tag{7-36}$$

where $\mathbf{x}(kT) = n$ vector (state vector)
$y(kT) =$ scalar (output signal)
$\mathbf{G} = n \times n$ nonsingular matrix
$\mathbf{C} = 1 \times n$ matrix
$T =$ sampling period

Before we present theorems on observability, we shall discuss some preliminary material. Notice that the output signal $y(kT)$ is given by

$$y(kT) = \mathbf{C}\mathbf{x}(kT) = (\mathbf{C}^*, \mathbf{x}(kT)) \qquad (k = 0, 1, 2, \ldots)$$

where $*$ denotes the conjugate transpose. $\mathbf{x}(kT)$ and $\mathbf{x}(0)$ are related by $\mathbf{x}(kT) = \mathbf{G}^k\,\mathbf{x}(0)$. Therefore, we can write $y(kT)$ as

$$y(kT) = (\mathbf{C}^*, \mathbf{G}^k\,\mathbf{x}(0)) = ((\mathbf{G}^*)^k\,\mathbf{C}^*, \mathbf{x}(0))$$

Let us put

$$(\mathbf{G}^*)^k\,\mathbf{C}^* = \boldsymbol{\varphi}_k^*$$

Note that the $\boldsymbol{\varphi}_k^*$ are column vectors. If this system is completely observable, then the initial state vector $\mathbf{x}(0)$ can be determined from n observations of the output.

Suppose we take n observations $y(T)$, $y(2T)$, $\ldots$, $y(nT)$, then n components of $\mathbf{x}(0)$ can be determined by solving the following n simultaneous equations for $x_1(0), x_2(0), \ldots, x_n(0)$:

$$y(kT) = (\boldsymbol{\varphi}_k^*, \mathbf{x}(0)) \qquad (k = 1, 2, \ldots, n)$$

or

$$\begin{aligned} \varphi_{11}x_1(0) + \varphi_{12}x_2(0) + \cdots + \varphi_{1n}x_n(0) &= y(T) \\ \varphi_{21}x_1(0) + \varphi_{22}x_2(0) + \cdots + \varphi_{2n}x_n(0) &= y(2T) \\ \cdots \qquad\quad \cdots \qquad\quad \cdots \qquad\quad & \\ \varphi_{n1}x_1(0) + \varphi_{n2}x_2(0) + \cdots + \varphi_{nn}x_n(0) &= y(nT) \end{aligned} \tag{7-37}$$

where $\varphi_{k1}, \varphi_{k2}, \ldots, \varphi_{kn}$ are n components of the row vector $\boldsymbol{\varphi}_k$. [Notice that if we take more than n observations $y(T)$, $y(2T)$, $\ldots$, we can obtain more than n equations of the form of Eq. (7–37). The number of linearly independent equations, however, is at most n.]

We shall next consider a necessary and sufficient condition that Eq. (7–37) can be solved for $x_1(0), x_2(0), \ldots, x_n(0)$. Equations (7–37) can be rewritten as

$$\boldsymbol{\Phi}\mathbf{x} = \mathbf{y} \tag{7-38}$$

where

$$\boldsymbol{\Phi} = \begin{bmatrix} \varphi_{11} & \varphi_{12} & \cdots & \varphi_{1n} \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{2n} \\ \vdots & \vdots & & \vdots \\ \varphi_{n1} & \varphi_{n2} & \cdots & \varphi_{nn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1(0) \\ x_2(0) \\ \vdots \\ x_n(0) \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y(T) \\ y(2T) \\ \vdots \\ y(nT) \end{bmatrix}$$

Theorem 7-12 A system of linear equations defined by Eq. (7–38) possesses a solution if and only if the matrix $\boldsymbol{\Phi}$ and the augumented matrix $[\boldsymbol{\Phi} \mid \mathbf{y}]$ where

$$[\boldsymbol{\Phi} \mid \mathbf{y}] = \begin{bmatrix} \varphi_{11} & \varphi_{12} & \cdots & \varphi_{1n} & y(T) \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{2n} & y(2T) \\ \vdots & \vdots & & \vdots & \vdots \\ \varphi_{n1} & \varphi_{n2} & \cdots & \varphi_{nn} & y(nT) \end{bmatrix}$$

are of the same rank. Moreover, the solution is unique if and only if rank $\boldsymbol{\Phi}$ = rank $[\boldsymbol{\Phi} \mid \mathbf{y}] = n$.

Proof Suppose that rank $\boldsymbol{\Phi}$ = rank $[\boldsymbol{\Phi} \mid \mathbf{y}] = r$ $(1 \leqslant r \leqslant n)$. We can assume, without loss of generality, that the first r columns of $\boldsymbol{\Phi}$ are linearly independent. Then, the first r columns of $[\boldsymbol{\Phi} \mid \mathbf{y}]$ are also linearly independent. Since the rank of $[\boldsymbol{\Phi} \mid \mathbf{y}]$ is assumed to be r, $\mathbf{y}$ must be a linear combination of the first r columns of $\boldsymbol{\Phi}$, or

$$\mathbf{y} = c_1 \begin{bmatrix} \varphi_{11} \\ \varphi_{21} \\ \vdots \\ \varphi_{n1} \end{bmatrix} + c_2 \begin{bmatrix} \varphi_{12} \\ \varphi_{22} \\ \vdots \\ \varphi_{n2} \end{bmatrix} + \cdots + c_r \begin{bmatrix} \varphi_{1r} \\ \varphi_{2r} \\ \vdots \\ \varphi_{nr} \end{bmatrix}$$

By choosing

$$\mathbf{x} = \begin{bmatrix} x_1(0) \\ x_2(0) \\ \vdots \\ x_r(0) \\ x_{r+1}(0) \\ \vdots \\ x_n(0) \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_r \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Eq. (7–38) can be satisfied. If $r = n$, the vector $\mathbf{x}$ which satisfies Eq. (7–38) is unique.

Conversely, if Eq. (7–38) possesses a solution, then we can write Eq. (7–38) as follows:

$$x_1(0)\begin{bmatrix}\varphi_{11}\\ \varphi_{21}\\ \vdots\\ \varphi_{n1}\end{bmatrix}+x_2(0)\begin{bmatrix}\varphi_{12}\\ \varphi_{22}\\ \vdots\\ \varphi_{n2}\end{bmatrix}+\cdots+x_n(0)\begin{bmatrix}\varphi_{1n}\\ \varphi_{2n}\\ \vdots\\ \varphi_{nn}\end{bmatrix}=\begin{bmatrix}y(T)\\ y(2T)\\ \vdots\\ y(nT)\end{bmatrix}$$

Clearly $\mathbf{y}$ is a linear combination of the columns of Φ. Hence

$$\text{rank } \Phi \geqslant \text{rank } [\Phi \mid \mathbf{y}]$$

Since, however, Φ is a submatrix of $[\Phi \mid \mathbf{y}]$ we have

$$\text{rank } \Phi \leqslant \text{rank } [\Phi \mid \mathbf{y}]$$

Hence,

$$\text{rank } \Phi = \text{rank } [\Phi \mid \mathbf{y}]$$

Moreover, if Eq. (7–38) possesses a unique solution, then the column vectors $\Phi_1, \Phi_2, \ldots, \Phi_n$ where $[\Phi_1 \mid \Phi_2 \mid \cdots \mid \Phi_n] = \Phi$ are linearly independent, or rank Φ is equal to n. Hence the theorem.

Notice that if $\mathbf{y} = \mathbf{0}$, then Eq. (7–38) becomes a homogeneous system, or $\Phi\mathbf{x} = \mathbf{0}$. If the rank of Φ is n (namely, if Φ is nonsingular) then the homogeneous system possesses a unique solution $\mathbf{x} = \mathbf{0}$, the so-called trivial solution. If, however, the rank of Φ is less than n (namely if Φ is singular), then the homogeneous system possesses nontrivial solutions.

A system of equations is said to be *consistent* if it has at least one solution. Otherwise it is said to be inconsistent. In the system of Eq. (7–38), the system is inconsistent if rank $\Phi <$ rank $[\Phi \mid \mathbf{y}]$. (The system possesses no solution.) If rank $\Phi =$ rank $[\Phi \mid \mathbf{y}] < n$, then the system is consistent but has infinitely many solutions.

Example 7-6 Consider the system $\Phi\mathbf{x} = \mathbf{y}$ where

$$\Phi = \begin{bmatrix}1 & 2 & 3\\ 1 & 4 & 6\\ 2 & 2 & 3\end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix}3\\ 4\\ 5\end{bmatrix}$$

The augmented matrix $[\Phi \mid \mathbf{y}]$ is given by

$$[\Phi \mid \mathbf{y}] = \begin{bmatrix}1 & 2 & 3 & 3\\ 1 & 4 & 6 & 4\\ 2 & 2 & 3 & 5\end{bmatrix}$$

Simple computation reveals that the rank of Φ is 2 and that of $[\Phi \mid \mathbf{y}]$ is also 2. Hence the system is consistent. But the system possesses infinitely many solutions. In fact,

$$\mathbf{x} = \begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix} = \begin{bmatrix}2\\ \frac{1}{2}(1-3a)\\ a\end{bmatrix}$$

where a is an arbitrary constant.

Now we are ready to present the following theorems on observability.

Theorem 7-13 The system described by Eq. (7–36), rewritten

$$\mathbf{x}((k+1)T) = \mathbf{G}\mathbf{x}(kT)$$
$$y(kT) = \mathbf{C}\mathbf{x}(kT)$$

where
$\mathbf{x}(kT) = n$ vector (state vector)
$y(kT) =$ scalar (output signal)
$\mathbf{G} = n \times n$ nonsingular matrix
$\mathbf{C} = 1 \times n$ matrix
$T =$ sampling period

is completely observable if and only if the composite $n \times n$ matrix $\mathbf{P}$ where

$$\mathbf{P} = [\mathbf{C}^* \mid \mathbf{G}^*\mathbf{C}^* \mid \cdots \mid (\mathbf{G}^*)^{n-1}\mathbf{C}^*]$$

is of rank n.

Proof $y(kT)$ can be written as

$$y(kT) = (\boldsymbol{\varphi}_k^*, \mathbf{x}(0)) \qquad (k = 1, 2, \ldots, n) \tag{7–39}$$

where $\boldsymbol{\varphi}_k^* = (\mathbf{G}^*)^k\mathbf{C}^*$

[Notice that the $\boldsymbol{\varphi}_k^*$, where $k = n+1, n+2, \ldots$, are linear combinations of $\boldsymbol{\varphi}_1^*, \boldsymbol{\varphi}_2^*, \ldots, \boldsymbol{\varphi}_n^*$. Hence we need consider only n values of $y(kT)$.] From Theorem 7–12 it follows that Eq. (7–39) possesses a unique solution $\mathbf{x}(0)$ if and only if the matrix Φ where

$$\Phi = \begin{bmatrix}\varphi_{11} & \varphi_{12} & \cdots & \varphi_{1n}\\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{2n}\\ \vdots & \vdots & & \vdots\\ \varphi_{n1} & \varphi_{n2} & \cdots & \varphi_{nn}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\varphi}_1\\ \boldsymbol{\varphi}_2\\ \vdots\\ \boldsymbol{\varphi}_n\end{bmatrix}$$

is of rank n, or

$$\Phi^* = \begin{bmatrix}\bar{\varphi}_{11} & \bar{\varphi}_{21} & \cdots & \bar{\varphi}_{n1}\\ \bar{\varphi}_{12} & \bar{\varphi}_{22} & \cdots & \bar{\varphi}_{n2}\\ \vdots & \vdots & & \vdots\\ \bar{\varphi}_{1n} & \bar{\varphi}_{2n} & \cdots & \bar{\varphi}_{nn}\end{bmatrix}$$
$$= [\boldsymbol{\varphi}_1^* \mid \boldsymbol{\varphi}_2^* \mid \cdots \mid \boldsymbol{\varphi}_n^*]$$
$$= [\mathbf{G}^*\mathbf{C}^* \mid (\mathbf{G}^*)^2\mathbf{C}^* \mid \cdots \mid (\mathbf{G}^*)^n\mathbf{C}^*]$$

is of rank n. Since $\mathbf{G}$ is nonsingular and

$$\Phi^* = [\mathbf{G}^*\mathbf{C}^* \mid \mathbf{G}^*\mathbf{G}^*\mathbf{C}^* \mid \cdots \mid \mathbf{G}^*(\mathbf{G}^*)^{n-1}\mathbf{C}^*]$$
$$= \mathbf{G}^*[\mathbf{C}^* \mid \mathbf{G}^*\mathbf{C}^* \mid \cdots \mid (\mathbf{G}^*)^{n-1}\mathbf{C}^*]$$

rank $\Phi^* = n$ implies that $\mathbf{P} = [\mathbf{C}^* \mid \mathbf{G}^*\mathbf{C}^* \mid \cdots \mid (\mathbf{G}^*)^{n-1}\mathbf{C}^*]$ is of rank n. Hence, the theorem.

Notice that this condition is equivalent to the condition that the composite $n \times n$ matrix $\mathbf{Q}$ where

$$\mathbf{Q} = \begin{bmatrix} \mathbf{C} \\ \hline \mathbf{CG} \\ \hline \vdots \\ \hline \mathbf{CG}^{n-1} \end{bmatrix}$$

is of rank n. Notice that $\mathbf{C}, \mathbf{CG}, \ldots, \mathbf{CG}^{n-1}$ are row vectors.

Since $\mathbf{G}$ is nonsingular the condition just obtained is also equivalent to the condition that the composite $n \times n$ matrix $\mathbf{S}$ where

$$\mathbf{S} = [(\mathbf{G}^*)^{-1}\mathbf{C}^* \mid (\mathbf{G}^*)^{-2}\mathbf{C}^* \mid \cdots \mid (\mathbf{G}^*)^{-n}\mathbf{C}^*]$$

is of rank n, or the composite $n \times n$ matrix $\mathbf{T}$ where

$$\mathbf{T} = \begin{bmatrix} \mathbf{CG}^{-1} \\ \hline \mathbf{CG}^{-2} \\ \hline \vdots \\ \hline \mathbf{CG}^{-n} \end{bmatrix}$$

is of rank n. Note that $\mathbf{CG}^{-1}, \mathbf{CG}^{-2}, \ldots, \mathbf{CG}^{-n}$ are row vectors.

Theorem 7-14 The system described by

$$\mathbf{x}((k+1)T) = \mathbf{G}\mathbf{x}(kT)$$
$$\mathbf{y}(kT) = \mathbf{C}\mathbf{x}(kT)$$

where
$\mathbf{x}(kT) = n$ vector (state vector)
$\mathbf{y}(kT) = m$ vector (output vector)
$\mathbf{G} = n \times n$ nonsingular matrix
$\mathbf{C} = m \times n$ matrix
$T =$ sampling period

is completely observable if and only if the composite $n \times mn$ matrix $\mathbf{P}$ where

$$\mathbf{P} = [\mathbf{C}^* \mid \mathbf{G}^*\mathbf{C}^* \mid \cdots \mid (\mathbf{G}^*)^{n-1}\mathbf{C}^*]$$

is of rank n.

Proof The proof is similar to that of Theorem 7–13. (It is also quite similar to that of Theorem 7–16 presented later.) We leave the proof to the reader.

We shall next consider observability of continuous-time, linear time-invariant systems.

Theorem 7-15 The system described by

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$$
$$y = \mathbf{C}\mathbf{x}$$

where
$\mathbf{x} = n$ vector (state vector)
$y =$ scalar (output signal)
$\mathbf{A} = n \times n$ matrix
$\mathbf{C} = 1 \times n$ matrix

is completely observable if and only if the composite $n \times n$ matrix $\mathbf{P}$ where

$$\mathbf{P} = [\mathbf{C}^* \mid \mathbf{A}^*\mathbf{C}^* \mid \cdots \mid (\mathbf{A}^*)^{n-1}\mathbf{C}^*]$$

is of rank n.

Proof First, note that if the system is completely observable, then the exact value of $\mathbf{x}(0)$ is determinable from measurements of $\mathbf{y}(t)$ over a time duration $0 \leqslant t \leqslant T$. Since

$$\mathbf{x}(t) = e^{\mathbf{A}t}\,\mathbf{x}(0)$$

and $e^{\mathbf{A}t}$ can be expressed as

$$e^{\mathbf{A}t} = \sum_{k=0}^{p-1} \alpha_k(t)\mathbf{A}^k$$

where p is the degree of the minimal polynomial of $\mathbf{A}$, we obtain,

$$\mathbf{x}(t) = \sum_{k=0}^{p-1} \alpha_k(t)\mathbf{A}^k\mathbf{x}(0)$$

Hence $y(t)$ can be written as

$$y(t) = \mathbf{C}\mathbf{x}(t) = \sum_{k=0}^{p-1} \alpha_k(t)\mathbf{C}\mathbf{A}^k\mathbf{x}(0) = \sum_{k=0}^{p-1} \alpha_k(t)((\mathbf{A}^*)^k\mathbf{C}^*, \mathbf{x}(0))$$

Let us put

$$(\mathbf{A}^*)^k\,\mathbf{C}^* = \boldsymbol{\varphi}_k^*$$

Taking the inner product of $\alpha_j(t)$ and $y(t)$, we obtain

$$\begin{aligned}(\alpha_j(t), y(t)) &= \sum_{k=0}^{p-1} (\alpha_j(t), \alpha_k(t))(\boldsymbol{\varphi}_k^*, \mathbf{x}(0)) \\ &= (\alpha_j(t), \alpha_0(t))(\boldsymbol{\varphi}_0^*, \mathbf{x}(0)) \\ &\quad + (\alpha_j(t), \alpha_1(t))(\boldsymbol{\varphi}_1^*, \mathbf{x}(0)) \\ &\quad + \cdots \\ &\quad + (\alpha_j(t), \alpha_{p-1}(t))(\boldsymbol{\varphi}_{p-1}^*, \mathbf{x}(0)) \qquad (j = 0, 1, \ldots, p-1)\end{aligned} \tag{7–40}$$

where

$$(\alpha_j(t), y(t)) = \int_0^T \alpha_j^*(t)y(t)\,dt$$

$$(\alpha_j(t), \alpha_k(t)) = \int_0^T \alpha_j^*(t)\alpha_k(t)dt$$

Equations (7–40) are a set of p equations in p variables $(\varphi_0^*, \mathbf{x}(0))$, $(\varphi_1^*, \mathbf{x}(0))$, $\ldots$, $(\varphi_{p-1}^*, \mathbf{x}(0))$. Rewriting in the form of a vector matrix equation, we obtain

$$\begin{bmatrix} (\alpha_0, \alpha_0) & (\alpha_0, \alpha_1) & \cdots & (\alpha_0, \alpha_{p-1}) \\ (\alpha_1, \alpha_0) & (\alpha_1, \alpha_1) & \cdots & (\alpha_1, \alpha_{p-1}) \\ \vdots & \vdots & & \vdots \\ (\alpha_{p-1}, \alpha_0) & (\alpha_{p-1}, \alpha_1) & \cdots & (\alpha_{p-1}, \alpha_{p-1}) \end{bmatrix} \begin{bmatrix} (\varphi_0^*, \mathbf{x}(0)) \\ (\varphi_1^*, \mathbf{x}(0)) \\ \vdots \\ (\varphi_{p-1}^*, \mathbf{x}(0)) \end{bmatrix} = \begin{bmatrix} (\alpha_0, y) \\ (\alpha_1, y) \\ \vdots \\ (\alpha_{p-1}, y) \end{bmatrix} \tag{7–41}$$

where $\alpha_k = \alpha_k(t)$ $(k = 0, 1, \ldots, p-1)$ and $y = y(t)$. Since the $\alpha_k(t)$ are linearly independent of each other, the Gram determinant $|(\alpha_i, \alpha_j)|$ is positive. Hence Eq. (7–41) can be solved for $(\varphi_0^*, \mathbf{x}(0))$, $(\varphi_1^*, \mathbf{x}(0))$, $\ldots$, $(\varphi_{p-1}^*, \mathbf{x}(0))$, giving

$$\begin{aligned} (\varphi_0^*, \mathbf{x}(0)) &= \beta_0 \\ (\varphi_1^*, \mathbf{x}(0)) &= \beta_1 \\ &\cdots\cdots \\ (\varphi_{p-1}^*, \mathbf{x}(0)) &= \beta_{p-1} \end{aligned} \tag{7–42}$$

Equations (7–42) are p simultaneous linear equations in $x_1(0)$, $x_2(0), \ldots, x_n(0)$. From Theorem 7–12 it follows that Eq. (7–42) can be solved uniquely for $x_1(0)$, $x_2(0)$, $\ldots$, $x_n(0)$ if and only if $p = n$ and the rank of the following matrix:

$$[\varphi_0^* \mid \varphi_1^* \mid \cdots \mid \varphi_{n-1}^*] = [\mathbf{C}^* \mid \mathbf{A}^*\mathbf{C}^* \mid \cdots \mid (\mathbf{A}^*)^{n-1}\mathbf{C}^*]$$

is n. This completes the proof.

Theorem 7-16 The system described by

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} \\ \mathbf{y} &= \mathbf{C}\mathbf{x} \end{aligned}$$

where

$\mathbf{x} = n$ vector (state vector)
$\mathbf{y} = m$ vector (output vector) $(m \leqslant n)$
$\mathbf{A} = n \times n$ matrix
$\mathbf{C} = m \times n$ matrix

is completely observable if and only if the composite $n \times mn$ matrix $\mathbf{P}$ where

$$\mathbf{P} = [\mathbf{C}^* \mid \mathbf{A}^*\mathbf{C}^* \mid \cdots \mid (\mathbf{A}^*)^{n-1}\mathbf{C}^*]$$

is of rank n.

Proof We shall give only the outline of the proof and shall leave details to the reader, since it is quite similar to that of Theorem 7–15. In the present system the matrix $\mathbf{C}^*$ has m columns, or

$$\mathbf{C}^* = [\mathbf{C}_1^* \mid \mathbf{C}_2^* \mid \cdots \mid \mathbf{C}_m^*]$$

The output vector $\mathbf{y}(t)$ has m components, its ith component being

$$y_i(t) = \sum_{k=0}^{p-1} \alpha_k(t)((\mathbf{A}^*)^k\mathbf{C}_i^*, \mathbf{x}(0)) \qquad (1 \leqslant i \leqslant m)$$

where p is the degree of the minimal polynomial of $\mathbf{A}$. Notice that taking the inner product of $\alpha_j(t)$ and $y_i(t)$, we obtain pm equations in pm variables $(\mathbf{C}_i^*, \mathbf{x}(0))$, $(\mathbf{A}^*\mathbf{C}_i^*, \mathbf{x}(0)), \ldots, ((\mathbf{A}^*)^{p-1}\mathbf{C}_i^*, \mathbf{x}(0))$, $i = 1, 2, \ldots, m$. Noting that the $\alpha_j(t)$ are linearly independent, we can solve the set of pm equations in pm unknown variables, giving

$$((\mathbf{A}^*)^k\mathbf{C}_i^*, \mathbf{x}(0)) = \beta_{ki}$$

which can be rewritten as

$$\begin{aligned} \varphi_{011}x_1(0) + \varphi_{012}x_2(0) + \cdots + \varphi_{01n}x_n(0) &= \beta_{01} \\ \varphi_{021}x_1(0) + \varphi_{022}x_2(0) + \cdots + \varphi_{02n}x_n(0) &= \beta_{02} \\ \cdots\cdots \qquad \cdots\cdots \qquad \cdots\cdots & \\ \varphi_{p-1m_1}x_1(0) + \varphi_{p-1m_2}x_2(0) + \cdots + \varphi_{p-1mn}x_n(0) &= \beta_{p-1m} \end{aligned} \tag{7–43}$$

where $\varphi_{ki1}, \varphi_{ki2}, \ldots, \varphi_{kin}$ are n components of the row vector φ_{ki} and

$$\varphi_{ki}^* = (\mathbf{A}^*)^k\mathbf{C}_i^*$$

Equations (7–43) are a set of pm equations in n unknown variables $x_1(0)$, $x_2(0), \ldots, x_n(0)$. These equations are necessarily consistent. They have a unique solution if and only if the rank of the $pm \times n$ coefficient matrix

$$\Phi = \begin{bmatrix} \varphi_{011} & \varphi_{012} & \cdots & \varphi_{01n} \\ \varphi_{021} & \varphi_{022} & \cdots & \varphi_{02n} \\ \vdots & \vdots & & \vdots \\ \varphi_{p-1m_1} & \varphi_{p-1m_2} & \cdots & \varphi_{p-1mn} \end{bmatrix}$$

is n. The conjugate transpose of Φ can be written as

$$\begin{aligned} \Phi^* &= \begin{bmatrix} \bar{\varphi}_{011} & \bar{\varphi}_{021} & \cdots & \bar{\varphi}_{p-1m_1} \\ \bar{\varphi}_{012} & \bar{\varphi}_{022} & \cdots & \bar{\varphi}_{p-1m_2} \\ \vdots & \vdots & & \vdots \\ \bar{\varphi}_{01n} & \bar{\varphi}_{02n} & \cdots & \bar{\varphi}_{p-1mn} \end{bmatrix} \\ &= [\varphi_{01}^* \mid \varphi_{02}^* \mid \cdots \mid \varphi_{p-1m}^*] \\ &= [\mathbf{C}_1^* \mid \mathbf{C}_2^* \mid \cdots \mid (\mathbf{A}^*)^{p-1}\mathbf{C}_m^*] \end{aligned}$$

Hence, the initial values $x_1(0)$, $x_2(0), \ldots, x_n(0)$ can be determined uniquely if and only if the rank of the following matrix:

$$\Phi^* = [\mathbf{C}_1^* \mid \mathbf{C}_2^* \mid \cdots \mid (\mathbf{A}^*)^{p-1}\mathbf{C}_m^*]$$
$$= [\mathbf{C}^* \mid \mathbf{A}^*\mathbf{C}^* \mid \cdots \mid (\mathbf{A}^*)^{p-1}\mathbf{C}^*]$$

is n. If $p=n$, then this condition can be stated that the rank of $\mathbf{P}$ where

$$\mathbf{P} = [\mathbf{C}^* \mid \mathbf{A}^*\mathbf{C}^* \mid \cdots \mid (\mathbf{A}^*)^{n-1}\mathbf{C}^*]$$

is n. Notice that if $p < n$, then the columns of

$$(\mathbf{A}^*)^h\mathbf{C}^* \qquad (h = p, p+1, \ldots, n-1)$$

are linearly dependent on $\mathbf{C}_1^*, \mathbf{C}_2^*, \ldots, (\mathbf{A}^*)^{p-1}\mathbf{C}_m^*$. The condition of complete observability in the case of $p < n$ can still be stated that the rank of $\mathbf{P}$ is n. This completes the proof.

It is noted that if the rank of $\mathbf{P}$ is n, then there is a set of n linearly independent vectors among the nm columns of $\mathbf{P}$. These vectors form a basis of the state space of the system. Hence the initial state $\mathbf{x}(0)$ can be determined by measurements of $\mathbf{y}(t)$ for a finite time interval.

Example 7-7 Examine observability of the following system:

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix} = \begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ -6 & -11 & -6\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix} + \begin{bmatrix}0\\ 0\\ 1\end{bmatrix}[u]$$

$$y = [20 \quad 9 \quad 1]\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}$$

In this system

$$\mathbf{A} = \begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ -6 & -11 & -6\end{bmatrix} \quad \text{and} \quad \mathbf{C} = [20 \quad 9 \quad 1]$$

Thus,

$$\mathbf{C}^* = \mathbf{C}' = \begin{bmatrix}20\\ 9\\ 1\end{bmatrix}$$

$$\mathbf{A}^*\mathbf{C}^* = \mathbf{A}'\mathbf{C}' = \begin{bmatrix}0 & 0 & -6\\ 1 & 0 & -11\\ 0 & 1 & -6\end{bmatrix}\begin{bmatrix}20\\ 9\\ 1\end{bmatrix} = \begin{bmatrix}-6\\ 9\\ 3\end{bmatrix}$$

$$(\mathbf{A}^*)^2\mathbf{C}^* = (\mathbf{A}')^2\mathbf{C}' = \begin{bmatrix}0 & -6 & 36\\ 0 & -11 & 60\\ 1 & -6 & 25\end{bmatrix}\begin{bmatrix}20\\ 9\\ 1\end{bmatrix} = \begin{bmatrix}-18\\ -39\\ -9\end{bmatrix}$$

It can be shown that the three vectors $\mathbf{C}'$, $\mathbf{A}'\mathbf{C}'$, $(\mathbf{A}')^2\mathbf{C}'$ are linearly independent. Hence the rank of the matrix $[\mathbf{C}' \mid \mathbf{A}'\mathbf{C}' \mid (\mathbf{A}')^2\mathbf{C}']$ is 3. By Theorem 7–15 the system is completely observable.

Example 7-8 Examine observability of the following system:

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix} = \begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ -6 & -11 & -6\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix} + \begin{bmatrix}0\\ 0\\ 1\end{bmatrix}[u]$$

$$y = [4 \quad 5 \quad 1]\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}$$

In this system,

$$\mathbf{C}^* = \mathbf{C}' = \begin{bmatrix}4\\ 5\\ 1\end{bmatrix}$$

$$\mathbf{A}^*\mathbf{C}^* = \mathbf{A}'\mathbf{C}' = \begin{bmatrix}0 & 0 & -6\\ 1 & 0 & -11\\ 0 & 1 & -6\end{bmatrix}\begin{bmatrix}4\\ 5\\ 1\end{bmatrix} = \begin{bmatrix}-6\\ -7\\ -1\end{bmatrix}$$

$$(\mathbf{A}^*)^2\mathbf{C}^* = (\mathbf{A}')^2\mathbf{C}' = \begin{bmatrix}0 & -6 & 36\\ 0 & -11 & 60\\ 1 & -6 & 25\end{bmatrix}\begin{bmatrix}4\\ 5\\ 1\end{bmatrix} = \begin{bmatrix}6\\ 5\\ -1\end{bmatrix}$$

It can be seen that

$$(\mathbf{A}')^2\mathbf{C}' = (-6)\mathbf{C}' + (-5)\mathbf{A}'\mathbf{C}'$$

We find that the rank of the matrix

$$[\mathbf{C}' \mid \mathbf{A}'\mathbf{C}' \mid (\mathbf{A}')^2\mathbf{C}']$$

is two. The system is therefore not completely observable. This can be seen clearly in the block diagram of Fig. 7–3. The two factors $(s+1)$ cancel each other. Hence,

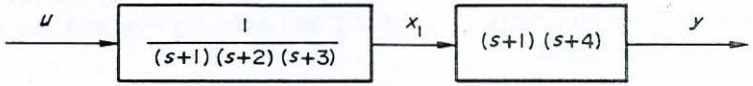


Figure 7-3 Control system of Example 7-8.

measurements of $y(t)$ do not yield estimation of all the initial values of the elements of $\mathbf{x}(t)$. In other words, there are nonzero initial states which cannot be determined from measurements of $y(t)$.

We can characterize the notion of observability in the s plane. Consider the following system:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}, \qquad \mathbf{x}(0) = \mathbf{x}_0$$
$$y = \mathbf{C}\mathbf{x} \tag{7–44}$$

where
$\mathbf{x} = n$ vector (state vector)
$y =$ scalar (output signal)
$\mathbf{A} = n \times n$ matrix
$\mathbf{C} = 1 \times n$ matrix

Taking Laplace transforms of Eq. (7–44) and simplifying we obtain

$$\begin{aligned} \mathbf{X}(s) &= (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{x}_0 \\ Y(s) &= \mathbf{C}\mathbf{X}(s) \end{aligned} \tag{7–45}$$

Elimination of $\mathbf{X}(s)$ from Eq. (7–45) results in

$$Y(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{x}_0$$

where

$$\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1} = \frac{1}{|s\mathbf{I} - \mathbf{A}|}[q_1(s) \quad q_2(s) \quad \cdots \quad q_n(s)]$$

and the $q_i(s)$ are polynomials in s.

If $\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}$ has a cancellation, namely the polynomials $q_1(s)$, $q_2(s)$, $\ldots$, $q_n(s)$, and $|s\mathbf{I} - \mathbf{A}|$ have a common factor, then this canceled mode cannot be observed in the output y. The following theorem relates complete observability and s plane representation of a linear time-invariant system.

Theorem 7-17 A necessary and sufficient condition that the system

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{B}u \\ y &= \mathbf{C}\mathbf{x} \end{aligned}$$

where

$\mathbf{x} = n$ vector (state vector)
$u =$ scalar (control signal)
$y =$ scalar (output signal)
$\mathbf{A} = n \times n$ matrix
$\mathbf{B} = n \times 1$ matrix
$\mathbf{C} = 1 \times n$ matrix

is completely observable is that $\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}$ has no cancellation.

Proof Notice that $\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}$ has a cancellation if and only if $[\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}]^*$ has a cancellation. Since

$$[\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}]^* = (s\mathbf{I} - \mathbf{A}^*)^{-1}\mathbf{C}^*$$

the proof of this theorem follows immediately from Theorems 7–8 and 7–15.

7-5 CONTROLLABILITY AND OBSERVABILITY OF LINEAR TIME-VARYING SYSTEMS (7-2, 7-3)

In this section we shall discuss controllability and observability of linear time-varying systems. We first present theorems on complete state controllability and complete output controllability of time-varying systems. Then we shall present a theorem on complete observability of time-varying systems.

Theorem 7-18 The linear time-varying system

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u} \tag{7–46}$$

where

$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{A}(t) = n \times n$ matrix, elements of which are piecewise continuous functions of time
$\mathbf{B}(t) = n \times r$ matrix, elements of which are piecewise continuous functions of time

is completely state controllable on $t_0 \leqslant t \leqslant t_1$ if and only if the Gram matrix $\mathbf{W}(t_0, t_1)$ is nonsingular, where

$$\mathbf{W}(t_0, t_1) = \int_{t_0}^{t_1} \Phi(t_0, t)\mathbf{B}(t)\mathbf{B}^*(t)\Phi^*(t_0, t)dt \tag{7–47}$$

and $\Phi(t, t_0)$ is the unique fundamental matrix satisfying

$$\frac{d}{dt}\Phi(t, t_0) = \mathbf{A}(t)\Phi(t, t_0), \qquad \Phi(t_0, t_0) = \mathbf{I}$$

(Since the Gram matrix is positive definite or positive semidefinite, if it is nonsingular, it is positive definite. Notice that the Gram matrix $\mathbf{W}(t_0, t_1)$ is Hermitian. If $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are real, then $\Phi(t_0, t)$ is real and $\mathbf{W}(t_0, t_1)$ becomes real symmetric.)

Proof The solution of Eq. (7–46) is given by

$$\mathbf{x}(t_1) = \Phi(t_1, t_0)\mathbf{x}(t_0) + \int_{t_0}^{t_1} \Phi(t_1, t)\mathbf{B}(t)\mathbf{u}(t)dt$$

Without loss of generality we may assume the final state to be the origin of the state space. Suppose that the Gram matrix $\mathbf{W}(t_0, t_1)$ is nonsingular. Then $\mathbf{W}^{-1}(t_0, t_1)$ exists. If we choose $\mathbf{u}(t)$ to be $\mathbf{u}_0(t)$ where

$$\mathbf{u}_0(t) = -\mathbf{B}^*(t)\Phi^*(t_0, t)\mathbf{W}^{-1}(t_0, t_1)\mathbf{x}(t_0) \tag{7–48}$$

then $\mathbf{x}(t_1) = \mathbf{0}$, since by the use of this particular $\mathbf{u}_0(t)$ we obtain

$$\begin{aligned} \mathbf{x}(t_1) &= \Phi(t_1, t_0)\mathbf{x}(t_0) \\ &\quad - \int_{t_0}^{t_1} \Phi(t_1, t)\mathbf{B}(t)\mathbf{B}^*(t)\Phi^*(t_0, t)dt\,\mathbf{W}^{-1}(t_0, t_1)\mathbf{x}(t_0) \\ &= \Phi(t_1, t_0)\mathbf{x}(t_0) \\ &\quad - \Phi(t_1, t_0)\int_{t_0}^{t_1} \Phi(t_0, t)\mathbf{B}(t)\mathbf{B}^*(t)\Phi^*(t_0, t)dt\,\mathbf{W}^{-1}(t_0, t_1)\mathbf{x}(t_0) \\ &= \Phi(t_1, t_0)\mathbf{x}(t_0) - \Phi(t_1, t_0)\mathbf{W}(t_0, t_1)\mathbf{W}^{-1}(t_0, t_1)\mathbf{x}(t_0) \\ &= \mathbf{0} \end{aligned}$$

where we have used the relation

$$\Phi(t_1, t) = \Phi(t_1, t_0)\Phi(t_0, t)$$

Hence if the Gram matrix $\mathbf{W}(t_0, t_1)$ is nonsingular, then any initial state $\mathbf{x}(t_0)$ can be transferred to the origin of the state space during the time interval $t_0 \leqslant t \leqslant t_1$.

Conversely, suppose the system is completely state controllable. Then we can prove that the Gram matrix $\mathbf{W}(t_0, t_1)$ is nonsingular. To prove this suppose $\mathbf{W}(t_0, t_1)$ is singular. Then there exists some state $\mathbf{x}(t_0) \neq \mathbf{0}$ such that $(\mathbf{x}(t_0), \mathbf{W}(t_0, t_1)\mathbf{x}(t_0)) = 0$. Define

$$\mathbf{u}_1(t) = -\mathbf{B}^*(t)\Phi^*(t_0, t)\mathbf{x}(t_0) \qquad \text{for } t_0 \leqslant t \leqslant t_1$$

Then

$$\begin{aligned}(\mathbf{x}(t_0), \mathbf{W}(t_0, t_1)\mathbf{x}(t_0)) &= \int_{t_0}^{t_1} (\mathbf{x}, \Phi\mathbf{B}\mathbf{B}^*\Phi^*\mathbf{x})dt \\ &= \int_{t_0}^{t_1} (\mathbf{B}^*\Phi^*\mathbf{x}, \mathbf{B}^*\Phi^*\mathbf{x})dt \\ &= \int_{t_0}^{t_1} (\mathbf{u}_1(t), \mathbf{u}_1(t))dt \\ &= \int_{t_0}^{t_1} \|\mathbf{u}_1(t)\|^2 dt = 0 \end{aligned} \tag{7–49}$$

Since $\mathbf{u}_1(t)$ is continuous in t, Eq. (7–49) implies that $\mathbf{u}_1(t) = \mathbf{0}$ for $t_0 \leqslant t \leqslant t_1$.

On the other hand, if the system is completely state controllable, then there exists a control vector $\mathbf{u}_0(t)$ for $t_0 \leqslant t \leqslant t_1$ defined by Eq. (7–48) which satisfies the condition

$$\mathbf{x}(t_0) = -\int_{t_0}^{t_1} \Phi(t_0, t)\mathbf{B}(t)\mathbf{u}_0(t)dt$$

Then

$$\begin{aligned}\|\mathbf{x}(t_0)\|^2 &= (\mathbf{x}(t_0), \mathbf{x}(t_0)) \\ &= (\mathbf{x}(t_0), -\int_{t_0}^{t_1} \Phi(t_0, t)\mathbf{B}(t)\mathbf{u}_0(t)dt) \\ &= -\int_{t_0}^{t_1} (\mathbf{x}(t_0), \Phi(t_0, t)\mathbf{B}(t)\mathbf{u}_0(t))dt \\ &= -\int_{t_0}^{t_1} (\mathbf{B}^*(t)\Phi^*(t_0, t)\mathbf{x}(t_0), \mathbf{u}_0(t))dt \\ &= \int_{t_0}^{t_1} (\mathbf{u}_1(t), \mathbf{u}_0(t))dt = 0 \end{aligned}$$

This is a contradiction, since we assumed that $\mathbf{x}(t_0) \neq \mathbf{0}$. Hence, $\mathbf{W}(t_0, t_1)$ must be nonsingular. This completes the proof.

Notice that the condition that the Gram matrix $\mathbf{W}(t_0, t_1)$ is nonsingular is equivalent to the statement that

$$\mathbf{B}^*(t)\Phi^*(t_1, t)\boldsymbol{\lambda} \neq \mathbf{0}$$

on $t_0 \leqslant t \leqslant t_1$ for each n vector $\boldsymbol{\lambda} \neq \mathbf{0}$ in the state space V. This can be seen as follows: The condition that the Gram matrix $\mathbf{W}(t_0, t_1)$ be nonsingular implies that the columns of $\mathbf{B}^*(t)\,\Phi^*(t_0, t)$ on $t_0 \leqslant t \leqslant t_1$ are linearly independent. This is equivalent to stating that

$$\mathbf{B}^*(t)\,\Phi^*(t_0, t)\,\boldsymbol{\lambda}_1 \neq \mathbf{0}$$

on $t_0 \leqslant t \leqslant t_1$ for each n vector $\boldsymbol{\lambda}_1 \neq \mathbf{0}$ in the state space. Since $\Phi(t_1, t_0)$ is nonsingular, let us define

$$\boldsymbol{\lambda} = [\Phi^*(t_1, t_0)]^{-1}\boldsymbol{\lambda}_1 \qquad \text{or} \qquad \boldsymbol{\lambda}_1 = \Phi^*(t_1, t_0)\boldsymbol{\lambda}$$

Since

$$\Phi(t_1, t) = \Phi(t_1, t_0)\Phi(t_0, t)$$

we obtain

$$\begin{aligned}\mathbf{B}^*(t)\Phi^*(t_0, t)\boldsymbol{\lambda}_1 &= \mathbf{B}^*(t)\Phi^*(t_0, t)\Phi^*(t_1, t_0)\boldsymbol{\lambda} \\ &= \mathbf{B}^*(t)\Phi^*(t_1, t)\boldsymbol{\lambda}\end{aligned}$$

Thus the condition that the Gram matrix $\mathbf{W}(t_0, t_1)$ is nonsingular is equivalent to the statement that

$$\mathbf{B}^*(t)\,\Phi^*(t_1, t)\,\boldsymbol{\lambda} \neq \mathbf{0}$$

on $t_0 \leqslant t \leqslant t_1$ for each $\boldsymbol{\lambda} \neq \mathbf{0}$ in the state space V. (See Prob. A-7-17.)

We shall next present a theorem on complete output controllability of linear time-varying systems.

Theorem 7-19 The linear time-varying system

$$\begin{aligned}\dot{\mathbf{x}} &= \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u} \\ \mathbf{y} &= \mathbf{C}(t)\mathbf{x}\end{aligned}$$

where
$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{y} = m$ vector (output vector) $(m \leqslant n)$
$\mathbf{A}(t) = n \times n$ matrix, elements of which are piecewise continuous functions of time
$\mathbf{B}(t) = n \times r$ matrix, elements of which are piecewise continuous functions of time
$\mathbf{C}(t) = m \times n$ matrix, elements of which are piecewise continuous functions of time

is completely output controllable on $t_0 \leqslant t \leqslant t_1$ if and only if the Gram matrix $\mathbf{V}(t_0, t_1)$ is nonsingular, where

$$\mathbf{V}(t_0, t_1) = \int_{t_0}^{t_1} \mathbf{C}(t_1)\Phi(t_1, t)\mathbf{B}(t)\mathbf{B}^*(t)\Phi^*(t_1, t)\mathbf{C}^*(t_1)dt$$

and $\Phi(t, t_0)$ is the unique fundamental matrix satisfying

$$\frac{d}{dt}\Phi(t, t_0) = \mathbf{A}(t)\Phi(t, t_0), \qquad \Phi(t_0, t_0) = \mathbf{I}$$

Proof The proof is similar to that of Prob. A-7-17. We leave details to the reader.

From Theorems 7–18 and 7–19 we can immediately prove that complete state controllability on $t_0 \leqslant t \leqslant t_1$ implies complete output controllability on $t_0 \leqslant t \leqslant t_1$ if and only if the m rows of $\mathbf{C}(t_1)$ are linearly independent.

We shall next consider complete observability of time-varying systems. We shall define observability of time-varying systems as follows: A time-varying system is said to be *completely observable on the time interval* $t_0 \leqslant t \leqslant t_1$ if for given t_0 and t_1 there is no initial state $\mathbf{x}(t_0) \neq \mathbf{0}$ in the state space V such that $\mathbf{y}(t) \equiv \mathbf{0}$ on $t_0 \leqslant t \leqslant t_1$. The system is said to be *completely observable* if for any t_0 and some t_1 there is no initial state such that $\mathbf{y}(t) \equiv \mathbf{0}$ for all $t \geqslant t_0$. The system is said to be *totally observable* if there is no initial state at any time such that $\mathbf{y}(t) \equiv \mathbf{0}$ on a nonzero time interval.

Theorem 7-20 *(7–2)* The system described by

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}(t)\mathbf{x} \\ \mathbf{y} &= \mathbf{C}(t)\mathbf{x} \end{aligned} \tag{7–50}$$

where $\mathbf{x} = n$ vector (state vector)
$\mathbf{y} = m$ vector (output vector) $(m \leqslant n)$
$\mathbf{A}(t) = n \times n$ matrix, elements of which are piecewise continuous functions of time
$\mathbf{C}(t) = m \times n$ matrix, elements of which are piecewise continuous functions of time

is completely observable on the time interval $t_0 \leqslant t \leqslant t_1$ if and only if the Gram matrix

$$\mathbf{M}(t_0, t_1) = \int_{t_0}^{t_1} \Phi^*(t, t_0)\mathbf{C}^*(t)\mathbf{C}(t)\Phi(t, t_0)dt \tag{7–51}$$

is nonsingular. Here $\Phi(t, t_0)$ is the unique fundamental matrix satisfying

$$\frac{d}{dt}\Phi(t, t_0) = \mathbf{A}(t)\Phi(t, t_0), \qquad \Phi(t_0, t_0) = \mathbf{I}$$

For complete observability, the foregoing must hold for every t_0 and some finite $t_1 > t_0$.

Proof The solution of the first of Eq. (7–50) is given by

$$\mathbf{x}(t) = \Phi(t, t_0)\mathbf{x}(t_0)$$

Then $\mathbf{y}(t)$ is obtained as

$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) = \mathbf{C}(t)\Phi(t, t_0)\mathbf{x}(t_0) \tag{7–52}$$

Suppose the Gram matrix $\mathbf{M}(t_0, t_1)$ is nonsingular. Then $\mathbf{M}^{-1}(t_0, t_1)$ exists. Premultiplying Eq. (7–52) by $\Phi^*(t, t_0)\mathbf{C}^*(t)$ and integrating on

$$\int_{t_0}^{t_1} \Phi^*(t, t_0)\mathbf{C}^*(t)\mathbf{y}(t)dt$$

$$= \left[\int_{t_0}^{t_1} \Phi^*(t, t_0)\mathbf{C}^*(t)\mathbf{C}(t)\Phi(t, t_0)dt\right]\mathbf{x}(t_0)$$

Hence $\mathbf{x}(t_0)$ can be expressed as

$$\mathbf{x}(t_0) = \mathbf{M}^{-1}(t_0, t_1)\int_{t_0}^{t_1} \Phi^*(t, t_0)\mathbf{C}^*(t)\mathbf{y}(t)dt \tag{7–53}$$

Every initial state $\mathbf{x}(t_0)$ can be determined from $\mathbf{y}(t)$ using Eq. (7–53).

Conversely, suppose the system is completely observable but the Gram matrix $\mathbf{M}(t_0, t_1)$ is singular. From Eq. (7–51) and Eq. (7–52) we obtain

$$\int_{t_0}^{t_1} \mathbf{y}^*(t)\mathbf{y}(t)dt = \mathbf{x}^*(t_0)\mathbf{M}(t_0, t_1)\mathbf{x}(t_0)$$

Since by assumption the Gram matrix $\mathbf{M}(t_0, t_1)$ is singular, there exists a state $\mathbf{x}(t_0) \neq \mathbf{0}$ such that $\mathbf{x}^*(t_0)\mathbf{M}(t_0, t_1)\mathbf{x}(t_0) = 0$. For such $\mathbf{x}(t_0)$, $\mathbf{y}(t) \equiv \mathbf{0}$ for the time interval $t_0 \leqslant t \leqslant t_1$. This implies that $\mathbf{x}(t_0)$ cannot be determined from $\mathbf{y}(t)$. This contradicts the assumption that the system is completely observable. Hence the theorem.

Notice that the conditions for complete state controllability, complete output controllability, and complete observability for time-invariant systems can be immediately derived from the corresponding conditions for time-varying systems (refer to Prob. A-7-18 and Prob. B-7-8).

In concluding this section, note that the concept of state controllability is essential in solving the regulator problem where we are concerned with transferring the state to the origin starting from a given initial point in the state space. If the system is not state controllable, then initial states induced by disturbances not in the controllable subspace cannot be controlled. This means that such a system may have long-term transients, or even instability. In certain cases control problems may be formulated in terms of the output $\mathbf{y}$ rather than the state $\mathbf{x}$. In such cases it is essential to have output controllability, rather than state controllability. Finally, note that if the system is not completely observable, then not all states $\mathbf{x}(t_0)$ can be determined from the output $\mathbf{y}(t)$, $t \geqslant t_0$.

7-6 RELATION BETWEEN CONTROLLABILITY AND OBSERVABILITY

In this section we shall discuss the relation between controllability and observability of dynamic systems. The analogies apparent between controllability and observability can be expressed by the principle of duality due to Kalman.

Principle of duality (7–1) Consider the system S_1 defined by

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u} \\ \mathbf{y} &= \mathbf{C}(t)\mathbf{x} \end{aligned} \tag{7-54}$$

where $\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{y} = m$ vector (output vector) $(m \leqslant n)$
$\mathbf{A}(t) = n \times n$ matrix, elements of which are piecewise continuous functions of time
$\mathbf{B}(t) = n \times r$ matrix, elements of which are piecewise continuous functions of time
$\mathbf{C}(t) = m \times n$ matrix, elements of which are piecewise continuous functions of time

and the dual system S_2 defined by

$$\begin{aligned} \dot{\mathbf{z}} &= -\mathbf{A}^*(t)\mathbf{z} + \mathbf{C}^*(t)\mathbf{v} \\ \mathbf{n} &= \mathbf{B}^*(t)\mathbf{z} \end{aligned} \tag{7-55}$$

where $\mathbf{z} = n$ vector (state vector)
$\mathbf{v} = m$ vector (control vector)
$\mathbf{n} = r$ vector (output vector)
$\mathbf{A}^*(t)$ = conjugate transpose of $\mathbf{A}(t)$
$\mathbf{B}^*(t)$ = conjugate transpose of $\mathbf{B}(t)$
$\mathbf{C}^*(t)$ = conjugate transpose of $\mathbf{C}(t)$

The system S_2 is obtained by

(1) Replacing the free dynamic system by its adjoint system.
(2) Interchanging input and output constraints.

The principle of duality states that the system S_1 is completely state controllable (observable) if and only if the system S_2 is completely observable (state controllable).

To verify this principle let us write down necessary and sufficient conditions for complete state controllability and complete observability of systems S_1 and S_2. For the system S_1

(1) A necessary and sufficient condition for complete state controllability is that

$$\int_{t_0}^{t_1} \Phi(t_0, t)\mathbf{B}(t)\mathbf{B}^*(t)\Phi^*(t_0, t)dt$$

is nonsingular.

(2) A necessary and sufficient condition for complete observability is that

$$\int_{t_0}^{t_1} \Phi^*(t, t_0)\mathbf{C}^*(t)\mathbf{C}(t)\Phi(t, t_0)\,dt$$

is nonsingular.

For the system S_2,

(1) A necessary and sufficient condition for complete state controllability is that

$$\int_{t_1}^{t_0} \Psi(t_0, t)\mathbf{C}^*(t)\mathbf{C}(t)\Psi^*(t_0, t)dt$$

is nonsingular.

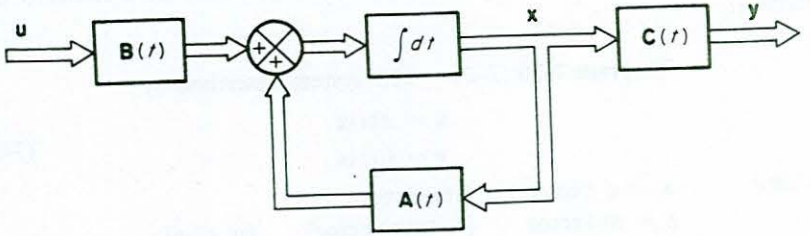
(2) A necessary and sufficient condition for complete observability is that

$$\int_{t_1}^{t_0} \Psi^*(t, t_0)\mathbf{B}(t)\mathbf{B}^*(t)\Psi(t, t_0)dt$$

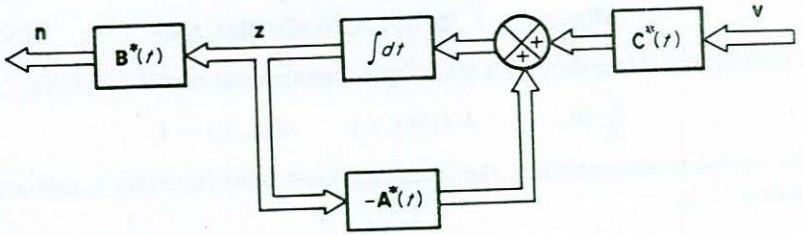
is nonsingular.

Notice that since $\Phi(t, t_0)$ and $\Psi(t, t_0)$ satisfy, respectively,

$$\frac{d}{dt}\Phi(t, t_0) = \mathbf{A}(t)\Phi(t, t_0), \qquad \Phi(t_0, t_0) = \mathbf{I}$$



System S_1



System S_2

Figure 7-4 System S_1 defined by Eq. (7-54) and its dual system S_2 defined by Eq. (7-55).

and

$$\frac{d}{dt}\Psi(t, t_0) = -\mathbf{A}^*(t)\Psi(t, t_0), \qquad \Psi(t_0, t_0) = \mathbf{I}$$

$\Phi(t_0, t)$ and $\Psi(t, t_0)$ are related by

$$\Phi(t_0, t) = \Psi^*(t, t_0)$$

By comparing the conditions, the truth of this principle is apparent. Figure 7–4 shows the system S_1 defined by Eq. (7–54) and its dual system S_2 defined by Eq. (7–55).

By the principle of duality the concept of observability may be dismissed by defining it as the dual of state controllability. The important point to note is that by the principle of duality a system constrained at the output can be converted into a system constrained at the input. If a system is completely state controllable (observable), then its dual system is completely observable (state controllable). Hence the solution of the minimal time observation problem becomes identical with the solution of the minimal time control problem for the dual system.

Note that the property of complete state controllability and the property of complete observability are invariant under linear nonsingular transformation.

Impulse response function Consider a system described by

$$\frac{X(s)}{U(s)} = \frac{(s+1)}{(s+1)(s+2)(s+3)} \tag{7–56}$$

This system is not completely state controllable. Referring to Example 4–4, the simplified state space representation of this system is given by Eq. (4–43), or

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} [u]$$

and x is given by $x = y_1 + y_2 - y_3$. The impulse response function $h(t)$ of the system defined by Eq. (7–56) is obtained as follows: Since for $u(t) = \delta(t)$ and $\mathbf{y}(0) = \mathbf{0}$

$$\mathbf{y}(t) = e^{\mathbf{A}t}\mathbf{B} = \begin{bmatrix} e^{-t} & 0 & 0 \\ 0 & e^{-2t} & 0 \\ 0 & 0 & e^{-3t} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ e^{-2t} \\ e^{-3t} \end{bmatrix}$$

the impulse response function $h(t)$ is given by

$$h(t) = e^{-2t} - e^{-3t}$$

Clearly, this impulse response function is the same as that of the system described by

$$\frac{X(s)}{U(s)} = \frac{1}{(s+2)(s+3)} \tag{7–57}$$

This system described by Eq. (7–57) is completely state controllable. Clearly, cancellation of the common terms in the transfer function defined by Eq. (7–56) results in the irreducible transfer function defined by Eq. (7–57). Note that the impulse response function does not represent the linear system completely, but represents only that part which is completely state controllable and completely observable.[1]

Let us next consider realization of a system having the impulse response function $h(t, \tau)$. We wish to realize the system with least order among a number of linear systems which possess $h(t, \tau)$ as the impulse response function.

Suppose we try to realize the impulse response function $h(t, \tau)$ by an nth-order system of the form given by

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)u \\ y &= \mathbf{C}(t)\mathbf{x} \end{aligned} \tag{7–58}$$

where
$\mathbf{x} = n$ vector (state vector)
$u =$ scalar (control signal)
$y =$ scalar (output signal)
$\mathbf{A}(t) = n \times n$ matrix
$\mathbf{B}(t) = n \times 1$ matrix
$\mathbf{C}(t) = 1 \times n$ matrix

The elements of $\mathbf{A}(t)$, $\mathbf{B}(t)$, and $\mathbf{C}(t)$ are assumed to be piecewise continuous functions of time. Note that impulse response function $h(t, \tau)$ can be realized by an nth-order system of the form given by Eq. (7–58) if and only if $h(t, \tau)$ is expressible as

$$h(t, \tau) = \mathbf{P}(t)\mathbf{Q}(\tau)$$

where $\mathbf{P}(t)$ and $\mathbf{Q}(\tau)$ are $1 \times n$ matrix and $n \times 1$ matrix, respectively, and the elements of $\mathbf{P}(t)$ and $\mathbf{Q}(t)$ are continuous functions of time t. To verify this statement note that $\Phi(t, \tau)$ can be written as

$$\Phi(t, \tau) = \Phi(t, t_0)\Phi(t_0, \tau)$$

---

[1] By means of a suitable coordinate transformation the linear dynamic system may be decomposed into the following four mutually exclusive subsystems:

Subsystem A: controllable but not observable
Subsystem B: controllable and observable
Subsystem C: neither controllable nor observable
Subsystem D: not controllable but observable

For detailed discussions of canonical structure of linear dynamic systems see, for example, (7–6, 7–7).

and

$$h(t,\tau) = 0 \qquad t < \tau$$
$$h(t,\tau) = \mathbf{C}(t)\Phi(t,\tau)\mathbf{B}(\tau)$$
$$= \mathbf{C}(t)\Phi(t,t_0)\Phi(t_0,\tau)\mathbf{B}(\tau), \qquad t \geqslant \tau$$
$$= \mathbf{P}(t)\mathbf{Q}(\tau)$$

where $\mathbf{P}(t) = \mathbf{C}(t)\Phi(t,t_0) = 1 \times n$ matrix

$\mathbf{Q}(\tau) = \Phi(t_0,\tau)\mathbf{B}(\tau) = n \times 1$ matrix

If the linear system having $h(t,\tau)$ as the impulse response function does not include a lower-order subsystem which possesses the same impulse response function, then the system is of least order.

Realization of $h(t,\tau)$ is irreducible (if the system is time-invariant, the system transfer function has no cancellation) if and only if the realized system is completely state controllable and completely observable. In other words realization of $h(t,\tau)$ by a completely state controllable and completely observable system yields least-order n.

Concluding comments In concluding this chapter we list reasons why concepts of controllability and observability are important in control theory.

1. Concepts of controllability and observability give a complete solution to the optimal regulator problem. The linear independence conditions for controllability and observability are of fundamental importance in the analysis and synthesis of multivariable control systems. Familiarity with concepts of controllability and observability enables one to formulate meaningful control problems and to make the necessary assumptions, if any, at the outset of the analysis.
2. Controllability and observability are dual. Therefore, Wiener's filtering problem can be considered as dual to the optimal regulator problem.
3. Concepts of controllability and observability clarify the relationship between the linear dynamic system and the impulse response function (or the transfer function).

EXERCISE PROBLEMS AND SOLUTIONS

Problem A-7-1 Consider the system

$$\overset{(n)}{x} + a_1\overset{(n-1)}{x} + \cdots + a_{n-1}\dot{x} + a_n x = u \tag{7-59}$$

and the adjoint system

$$(-1)^n\overset{(n)}{y} + (-1)^{n-1}\bar{a}_1\overset{(n-1)}{y} + \cdots + (-1)\bar{a}_{n-1}\dot{y} + \bar{a}_n y = u \tag{7-60}$$

where the bar denotes the complex conjugate. Let $h_1(t)$ and $h_2(t)$ be the impulse response functions of the systems described by Eq. (7-59) and Eq. (7-60), respectively. Show that

$$\bar{h}_1(t) = -h_2(-t)$$

Solution Taking the Laplace transform of Eq. (7-60) with

$$\overset{(n-1)}{y}(0) = \overset{(n-2)}{y}(0) = \ldots = \dot{y}(0) = y(0) = 0$$

the result is

$$[(-s)^n + \bar{a}_1(-s)^{n-1} + \cdots + \bar{a}_{n-1}(-s) + \bar{a}_n]Y(s) = U(s) \tag{7-61}$$

The impulse response function $h_2(\tau)$ is then obtained from Eq. (7-61) as

$$h_2(t) = \mathscr{L}^{-1}\left[\frac{1}{(-s)^n + \bar{a}_1(-s)^{n-1} + \cdots + \bar{a}_{n-1}(-s) + \bar{a}_n}\right], \qquad t > 0$$
$$h_2(t) = 0 \qquad t < 0 \tag{7-62}$$

The Laplace transform of Eq. (7-59) with zero initial condition is

$$(s^n + a_1 s^{n-1} + \cdots + a_{n-1}s + a_n)X(s) = U(s)$$

The impulse response function $h_1(t)$ is then equal to the inverse Laplace transform of $X(s)$ with $U(s) = 1$. Thus

$$h_1(t) = \mathscr{L}^{-1}\left[\frac{1}{s^n + a_1 s^{n-1} + \cdots + a_{n-1}s + a_n}\right], \qquad t > 0$$
$$h_1(t) = 0 \qquad t < 0$$

The complex conjugate of $h_1(t)$; namely, $\bar{h}_1(t)$, is

$$\bar{h}_1(t) = \mathscr{L}^{-1}\left[\frac{1}{s^n + \bar{a}_1 s^{n-1} + \cdots + \bar{a}_{n-1}s + \bar{a}_n}\right], \qquad t > 0 \tag{7-63}$$

Thus comparing Eq. (7-62) with Eq. (7-63) and noting that

$$\mathscr{L}^{-1}\left[\frac{1}{s+a}\right] = e^{-at}$$

and

$$\mathscr{L}^{-1}\left[\frac{1}{(-s)+a}\right] = -e^{at}$$

we obtain

$$\bar{h}_1(t) = -h_2(-t)$$

Problem A-7-2 Consider a system described by

$$\dot{\mathbf{X}} = \mathbf{A}(t)\mathbf{X} \tag{7-64}$$

and a system described by

$$\dot{\mathbf{Y}} = -\mathbf{Y}\mathbf{A}(t) \tag{7-65}$$

Show that if $\mathbf{Y}$ is a nonsingular solution of Eq. (7–65), then $\mathbf{Y}^{-1}\mathbf{C}$ where $\mathbf{C}$ is a constant matrix is a solution of Eq. (7–64).

Solution From Eq. (7–64) and Eq. (7–65)

$$\mathbf{Y}\dot{\mathbf{X}} + \dot{\mathbf{Y}}\mathbf{X} = \mathbf{0} \tag{7–66}$$

Hence,

$$\mathbf{Y}\mathbf{X} = \mathbf{C}$$

where $\mathbf{C}$ is a constant matrix. Since $\mathbf{Y}$ is nonsingular, $\mathbf{Y}^{-1}$ exists and

$$\mathbf{X} = \mathbf{Y}^{-1}\mathbf{C} \tag{7–67}$$

Clearly, Eq. (7–67) satisfies Eq. (7–66), or

$$\mathbf{Y}\dot{\mathbf{X}} = \mathbf{Y}\mathbf{A}(t)\mathbf{X}$$

which is equivalent to

$$\dot{\mathbf{X}} = \mathbf{A}(t)\mathbf{X}$$

Thus, $\mathbf{Y}^{-1}\mathbf{C}$ is a solution of Eq. (7–64).

Problem A-7-3 Prove that if n vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are orthogonal to each other, then these vectors are linearly independent.

Solution Let us compute the Gram determinant.

$$|(\mathbf{x}_i, \mathbf{x}_j)| = \begin{vmatrix} (\mathbf{x}_1, \mathbf{x}_1) & (\mathbf{x}_1, \mathbf{x}_2) & \cdots & (\mathbf{x}_1, \mathbf{x}_n) \\ (\mathbf{x}_2, \mathbf{x}_1) & (\mathbf{x}_2, \mathbf{x}_2) & \cdots & (\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & & \vdots \\ (\mathbf{x}_n, \mathbf{x}_1) & (\mathbf{x}_n, \mathbf{x}_2) & \cdots & (\mathbf{x}_n, \mathbf{x}_n) \end{vmatrix}$$

$$= \begin{vmatrix} (\mathbf{x}_1, \mathbf{x}_1) & 0 & \cdots & 0 \\ 0 & (\mathbf{x}_2, \mathbf{x}_2) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & (\mathbf{x}_n, \mathbf{x}_n) \end{vmatrix}$$

$$= (\mathbf{x}_1, \mathbf{x}_1)(\mathbf{x}_2, \mathbf{x}_2) \cdots (\mathbf{x}_n, \mathbf{x}_n) \neq 0$$

By Theorem 7–2, n vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are linearly independent.

Problem A-7-4 Consider the system

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

where $\mathbf{x}$ and $\mathbf{y}$ are n vectors and $\mathbf{A}$ is an $n \times n$ matrix. Prove that

$$((y_i, y_j)) = \bar{\mathbf{A}}((x_i, x_j))\mathbf{A}'$$

where $\bar{\mathbf{A}}$ is the complex conjugate of $\mathbf{A}$. Prove also that

$$|(y_i, y_j)| = |\mathbf{A}|^2 |(x_i, x_j)|$$

Solution Notice that

$$\bar{\mathbf{A}}((x_i, x_j))\mathbf{A}'$$

$$= \begin{bmatrix} \bar{a}_{11} & \cdots & \bar{a}_{1n} \\ \vdots & & \vdots \\ \bar{a}_{n1} & \cdots & \bar{a}_{nn} \end{bmatrix} \begin{bmatrix} (x_1, x_1) & \cdots & (x_1, x_n) \\ \vdots & & \vdots \\ (x_n, x_1) & \cdots & (x_n, x_n) \end{bmatrix} \begin{bmatrix} a_{11} & \cdots & a_{n1} \\ \vdots & & \vdots \\ a_{1n} & \cdots & a_{nn} \end{bmatrix}$$

$$= \begin{bmatrix} \bar{a}_{11} & \cdots & \bar{a}_{1n} \\ \vdots & & \vdots \\ \bar{a}_{n1} & \cdots & \bar{a}_{nn} \end{bmatrix} \begin{bmatrix} (x_1, y_1) & \cdots & (x_1, y_n) \\ \vdots & & \vdots \\ (x_n, y_1) & \cdots & (x_n, y_n) \end{bmatrix}$$

$$= \begin{bmatrix} (y_1, y_1) & \cdots & (y_1, y_n) \\ \vdots & & \vdots \\ (y_n, y_1) & \cdots & (y_n, y_n) \end{bmatrix} = ((y_i, y_j))$$

Taking determinants we obtain

$$|(y_i, y_j)| = |\bar{\mathbf{A}}||(x_i, y_j)||\mathbf{A}'| = |\mathbf{A}|^2 |(x_i, x_j)|$$

Problem A-7-5 (*Schmidt Orthonormalization*) Suppose that vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are a set of basis vectors of a state space V. Prove that it is always possible to find a set of orthogonal vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ as a basis of V.

Notice that this problem can be reduced to that of proving the existence of the coefficients $c_{11}, c_{21}, c_{22}, \ldots, c_{nn}$ such that

$$\begin{aligned} \mathbf{y}_1 &= c_{11}\mathbf{x}_1 \\ \mathbf{y}_2 &= c_{21}\mathbf{x}_1 + c_{22}\mathbf{x}_2 \\ &\cdots \quad \cdots \quad \cdots \\ \mathbf{y}_n &= c_{n1}\mathbf{x}_1 + c_{n2}\mathbf{x}_2 + \cdots + c_{nn}\mathbf{x}_n \end{aligned}$$

are orthonormal or $(\mathbf{y}_i, \mathbf{y}_j) = \delta_{ij}$.

Solution The following method for converting a given set of vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ into an orthonormal set of vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ is called the *Schmidt orthonormalization* method. In this method we begin with putting $\mathbf{x}_1 = \mathbf{z}_1$ and normalizing the vector $\mathbf{z}_1$ by dividing it by its norm $\|\mathbf{z}_1\|$ and setting $\mathbf{y}_1 = \mathbf{z}_1/\|\mathbf{z}_1\|$. We then define a vector $\mathbf{z}_2$, orthogonal to $\mathbf{y}_1$ by setting $\mathbf{z}_2 = \mathbf{x}_2 - (\mathbf{x}_2, \mathbf{y}_1)\mathbf{y}_1$. Then we normalize $\mathbf{z}_2$ by setting $\mathbf{y}_2 = \mathbf{z}_2/\|\mathbf{z}_2\|$. Next we define $\mathbf{z}_3$ orthogonal to $\mathbf{y}_1$ and $\mathbf{y}_2$ by setting $\mathbf{z}_3 = \mathbf{x}_3 - (\mathbf{x}_3, \mathbf{y}_1)\mathbf{y}_1 - (\mathbf{x}_3, \mathbf{y}_2)\mathbf{y}_2$ and then set $\mathbf{y}_3 = \mathbf{z}_3/\|\mathbf{z}_3\|$. The process is continued until we obtain $\mathbf{z}_n$ and $\mathbf{y}_n$. Rewritten $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$:

$$\begin{aligned} \mathbf{z}_1 &= \mathbf{x}_1, \qquad \mathbf{y}_1 = \frac{\mathbf{z}_1}{\|\mathbf{z}_1\|} \\ \mathbf{z}_2 &= \mathbf{x}_2 - (\mathbf{x}_2, \mathbf{y}_1)\mathbf{y}_1, \qquad \mathbf{y}_2 = \frac{\mathbf{z}_2}{\|\mathbf{z}_2\|} \\ &\cdots \quad \cdots \quad \cdots \\ \mathbf{z}_n &= \mathbf{x}_n - \sum_{k=1}^{n-1} (\mathbf{x}_n, \mathbf{y}_k)\mathbf{y}_k, \qquad \mathbf{y}_n = \frac{\mathbf{z}_n}{\|\mathbf{z}_n\|} \end{aligned} \tag{7–68}$$

Notice that $\|\mathbf{z}_k\| \neq 0$, since the $\mathbf{x}_i$ are linearly independent. The inner product of $\mathbf{y}_i$ and $\mathbf{y}_i$ is

$$(\mathbf{y}_i, \mathbf{y}_i) = \left(\frac{\mathbf{z}_i}{\|\mathbf{z}_i\|}, \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|}\right) = \left(\frac{(\mathbf{z}_i, \mathbf{z}_i)}{\|\mathbf{z}_i\|^2}\right) = 1$$

Hence the $\mathbf{y}_i$ are normal.

Notice that

$$(\mathbf{y}_2, \mathbf{y}_1) = \frac{1}{\|\mathbf{z}_2\|}(\mathbf{z}_2, \mathbf{y}_1)$$

$$= \frac{1}{\|\mathbf{z}_2\|}\Big[(\mathbf{x}_2, \mathbf{y}_1) - (\mathbf{x}_2, \mathbf{y}_1)(\mathbf{y}_1, \mathbf{y}_1)\Big] = 0$$

$$(\mathbf{y}_3, \mathbf{y}_1) = \frac{1}{\|\mathbf{z}_3\|}((\mathbf{x}_3 - (\mathbf{x}_3, \mathbf{y}_1)\mathbf{y}_1 - (\mathbf{x}_3, \mathbf{y}_2)\mathbf{y}_2), \mathbf{y}_1)$$

$$= \frac{1}{\|\mathbf{z}_3\|}\Big[(\mathbf{x}_3, \mathbf{y}_1) - (\mathbf{x}_3, \mathbf{y}_1)(\mathbf{y}_1, \mathbf{y}_1) - (\mathbf{x}_3, \mathbf{y}_2)(\mathbf{y}_2, \mathbf{y}_1)\Big] = 0$$

$$(\mathbf{y}_3, \mathbf{y}_2) = \frac{1}{\|\mathbf{z}_3\|}\Big[(\mathbf{x}_3, \mathbf{y}_2) - (\mathbf{x}_3, \mathbf{y}_1)(\mathbf{y}_1, \mathbf{y}_2) - (\mathbf{x}_3, \mathbf{y}_2)(\mathbf{y}_2, \mathbf{y}_2)\Big] = 0$$

Similarly, we can show that $\mathbf{y}_{m+1}$ is orthogonal to $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m$ $(m \leqslant n-1)$, since

$$(\mathbf{z}_{m+1}, \mathbf{y}_j) = (\mathbf{x}_{m+1} - \sum_{k=1}^{m}(\mathbf{x}_{m+1}, \mathbf{y}_k)\mathbf{y}_k, \mathbf{y}_j)$$

$$= (\mathbf{x}_{m+1}, \mathbf{y}_j) - \sum_{k=1}^{m}(\mathbf{x}_{m+1}, \mathbf{y}_k)(\mathbf{y}_k, \mathbf{y}_j)$$

$$= (\mathbf{x}_{m+1}, \mathbf{y}_j) - (\mathbf{x}_{m+1}, \mathbf{y}_j) = 0 \qquad (1 \leqslant j \leqslant m)$$

Therefore, the basis $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$ is orthonormal, or

$$(\mathbf{y}_i, \mathbf{y}_j) = \delta_{ij} \qquad (i, j = 1, 2, \ldots, n)$$

Equations (7–68) can be rewritten as follows:

$$\begin{aligned}
\mathbf{y}_1 &= c_{11}\mathbf{x}_1 \\
\mathbf{y}_2 &= c_{21}\mathbf{x}_1 + c_{22}\mathbf{x}_2 \\
&\cdots \quad \cdots \quad \cdots \\
\mathbf{y}_n &= c_{n1}\mathbf{x}_1 + c_{n2}\mathbf{x}_2 + \cdots + c_{nn}\mathbf{x}_n
\end{aligned}$$

The constants $c_{11}, c_{21}, \ldots, c_{nn}$ are determined from Eq. (7–68).

Problem A-7-6 Prove that Gram determinant $|(\mathbf{x}_i, \mathbf{x}_j)|$ is nonnegative.

Solution From Theorem 7–2, we have $|(\mathbf{x}_i, \mathbf{x}_j)| = 0$ if the $\mathbf{x}_i$ $(i = 1, 2, \ldots, n)$ are linearly dependent. If the $\mathbf{x}_i$ are linearly independent, then there exist constants c_{ij} such that

$$\begin{aligned}
\mathbf{y}_1 &= c_{11}\mathbf{x}_1 \\
\mathbf{y}_2 &= c_{21}\mathbf{x}_1 + c_{22}\mathbf{x}_2 \\
\mathbf{y}_3 &= c_{31}\mathbf{x}_1 + c_{32}\mathbf{x}_2 + c_{33}\mathbf{x}_3 \\
&\cdots \quad \cdots \quad \cdots \\
\mathbf{y}_n &= c_{n1}\mathbf{x}_1 + c_{n2}\mathbf{x}_2 + \cdots + c_{nn}\mathbf{x}_n
\end{aligned}$$

are orthonormal. Then from Prob. A-7-4, we obtain

$$|(\mathbf{x}_i, \mathbf{x}_j)||\mathbf{C}|^2 = |(\mathbf{y}_i, \mathbf{y}_j)| = |\delta_{ij}| = 1$$

where

$$|\mathbf{C}| = \begin{vmatrix} c_{11} & 0 & \cdots & 0 \\ c_{21} & c_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{vmatrix} = c_{11}c_{22} \cdots c_{nn}$$

Therefore, we obtain

$$|(\mathbf{x}_i, \mathbf{x}_j)| = \frac{1}{c_{11}^2}\frac{1}{c_{22}^2}\cdots\frac{1}{c_{nn}^2} > 0 \qquad \text{(if the } \mathbf{x}_i \text{ are linearly independent)}$$

$$|(\mathbf{x}_i, \mathbf{x}_j)| = 0 \qquad \text{(if the } \mathbf{x}_i \text{ are linearly dependent.)}$$

$|(\mathbf{x}_i, \mathbf{x}_j)|$ is thus nonnegative.

Problem A-7-7 Prove that any positive definite matrices and positive semidefinite matrices can be expressible as Gram matrices.

Solution Notice first that a Hermitian form $\mathbf{x}^*\mathbf{B}\mathbf{x}$ can be reduced to the diagonal Hermitian form by means of a suitable unitary transformation $\mathbf{x} = \mathbf{U}\mathbf{y}$, namely,

$$\mathbf{x}^*\mathbf{B}\mathbf{x} = \mathbf{y}^*\mathbf{U}^*\mathbf{B}\mathbf{U}\mathbf{y} = \lambda_1\bar{y}_1y_1 + \lambda_2\bar{y}_2y_2 + \cdots + \lambda_n\bar{y}_ny_n$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$, are eigenvalues of $\mathbf{B}$. If $\mathbf{B}$ is positive definite then $\lambda_1, \lambda_2, \ldots, \lambda_n$ are all positive real. If we choose

$$\mathbf{T} = \mathbf{U}\begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} & & & 0 \\ & \frac{1}{\sqrt{\lambda_2}} & & \\ & & \ddots & \\ 0 & & & \frac{1}{\sqrt{\lambda_n}} \end{bmatrix}$$

then the unitary transformation $\mathbf{x} = \mathbf{T}\mathbf{z}$ will transform $\mathbf{x}^*\mathbf{B}\mathbf{x}$ into the following form:

$$\mathbf{x}^*\mathbf{B}\mathbf{x} = \mathbf{z}^*\mathbf{T}^*\mathbf{B}\mathbf{T}\mathbf{z} = \bar{z}_1z_1 + \bar{z}_2z_2 + \cdots + \bar{z}_nz_n$$

Hence, if $\mathbf{B}$ is positive definite, then there exists a nonsingular matrix $\mathbf{T}$ such that

$$\mathbf{T}^*\mathbf{B}\mathbf{T} = \mathbf{I} \qquad \text{or} \qquad \mathbf{B} = (\mathbf{T}^*)^{-1}\mathbf{T}^{-1} = (\mathbf{T}^{-1})^*\mathbf{T}^{-1} = \mathbf{A}^*\mathbf{A}$$

where

$$\mathbf{A} = \mathbf{T}^{-1}$$

Any positive definite matrix can thus be represented as a Gram matrix $\mathbf{A}^*\mathbf{A}$.

Similarly, if $\mathbf{B}$ is positive semidefinite and has rank m, then there exists a nonsingular matrix $\mathbf{T}$ such that

$$\mathbf{T}^*\mathbf{B}\mathbf{T} = \left[\begin{array}{c|c} \mathbf{I}_m & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]$$

where $\mathbf{I}_m$ is the $m \times m$ identity matrix. Hence

$$\mathbf{B} = (\mathbf{T}^*)^{-1}\left[\begin{array}{c|c} \mathbf{I}_m & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]\mathbf{T}^{-1}$$

$$= (\mathbf{T}^{-1})^*\left[\begin{array}{c|c} \mathbf{I}_m & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]^*\left[\begin{array}{c|c} \mathbf{I}_m & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]\mathbf{T}^{-1}$$

$$= \mathbf{A}^*\mathbf{A}$$

where

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{I}_m & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]\mathbf{T}^{-1}$$

Thus any positive semidefinite matrix can be represented as a Gram matrix $\mathbf{A}^*\mathbf{A}$.

Problem A-7-8 Suppose that $\mathbf{G}$ is the Gram matrix of an $n \times n$ matrix $\mathbf{A}$. If $\mathbf{A}$ is of rank m ($m \leqslant n$), then prove that $\mathbf{G}$ is also of rank m.

Solution Since $\mathbf{G}$ is the Gram matrix of $\mathbf{A}$, we have

$$\mathbf{G} = \mathbf{A}^*\mathbf{A}$$

In Prob. A-7-7 we showed that

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{I}_m & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]\mathbf{T}^{-1}$$

where $\mathbf{I}_m$ is the $m \times m$ identity matrix and $\mathbf{T}$ is an $n \times n$ nonsingular matrix. Hence

$$\mathbf{G} = (\mathbf{T}^{-1})^*\left[\begin{array}{c|c} \mathbf{I}_m & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]\mathbf{T}^{-1} = \left[\begin{array}{c|c} \mathbf{B}_m & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]$$

where $\mathbf{B}_m$ is an $m \times m$ nonsingular matrix. The rank of $\mathbf{G}$ is therefore m.

Problem A-7-9 Consider a discrete-time system with a single control signal. The value of the control signal can take three values $+u_0$, 0, and $-u_0$. How many initial states can be transferred to the origin in n sampling periods?

Solution There are three possible values for u. Hence 3^n different combinations of $u(0)$, $u(1), \ldots, u(n)$ exist. Hence at most 3^n initial states can be transferred to the origin in n sampling periods.

Problem A-7-10 Given the state space equation

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$$

where

$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{A} = n \times n$ constant matrix
$\mathbf{B} = n \times r$ constant matrix

find the condition on $\mathbf{A}$ and $\mathbf{B}$ such that all modes of oscillation can be excited.

Solution Derivation of the condition on $\mathbf{A}$ and $\mathbf{B}$ for this problem can be accomplished using the derivation for complete state controllability. Suppose

that the system is oscillating in all modes. Then suppose that at some finite time t_1 it is required that all modes of oscillation have zero amplitude. Damping of the oscillations is accomplished by applying the control vector $\mathbf{u}$. Then to reduce all oscillations to zero at time t_1 it is necessary that $\mathbf{x}(t_1)$ be zero. It is obvious that all modes of oscillation must be excitable to do this. The problem can then be viewed as that of complete state controllability. Hence, the condition that all modes of oscillation can be excited is that n columns out of nr columns of

$$\mathbf{B}, \mathbf{AB}, \mathbf{A}^2\mathbf{B}, \ldots, \mathbf{A}^{n-1}\mathbf{B}$$

be linearly independent.

Problem A-7-11 An important thing to remember in discretizing continuous-time systems is that the introduction of sampling may impair controllability. A single-input–single-output system (continuous-time or discrete-time) is completely state controllable if and only if no cancellation of poles is possible in the transfer function. Notice that the cancellation may take place in passing from the continuous-time to the discrete-time case. A system which is completely state controllable in the absence of sampling remains completely state controllable after the introduction of sampling if and only if, for every eigenvalue of the characteristic equation,

$$\operatorname{Re}\lambda_i = \operatorname{Re}\lambda_j$$

implies

$$\operatorname{Im}[\lambda_i - \lambda_j] \neq \frac{2r\pi}{T}$$

where $r = \pm 1, \pm 2, \ldots$
$T =$ sampling period

(Unless the system contains complex poles, the cancellation will not occur in passing from the continuous-time to the discrete-time case.) Verify the foregoing statement by using the following systems:

(1) $$G(s) = \frac{s+a}{(s+a)^2 + b^2}$$

(2) $$G(s) = \frac{3}{(s^2+1)(s^2+4)}$$

Solution

(1) The z transform of

$$G(s) = \frac{s+a}{(s+a)^2 + b^2}$$

is given by

$$\mathcal{Z}[G] = \frac{e^{aT}z(e^{aT}z - \cos bT)}{e^{2aT}z^2 - 2(e^{aT}\cos bT)z + 1}$$

The two roots of the characteristic equation are $\lambda_1 = -a + jb$ and $\lambda_2 = -a - jb$. If $\operatorname{Im}[\lambda_1 - \lambda_2] = 2b = 2r\pi/T$ or $bT = r\pi$ (where $r =$ integer) the factor $(e^{aT}z - \cos r\pi)$ in both the numerator and the denominator will cancel. Hence for complete state controllability, we must have $T \neq (r\pi/b)$ (where $r = \pm 1, \pm 2, \ldots$).

(2) The z transform of

$$G(s) = \frac{3}{(s^2+1)(s^2+4)}$$

is given by

$$\mathcal{Z}[G] = \frac{z \sin T}{z^2 - 2z\cos T + 1} - \frac{z \sin 2T}{2(z^2 - 2z\cos 2T + 1)}$$

The four roots of the characteristic equation are $\lambda_1 = j$, $\lambda_2 = -j$, $\lambda_3 = 2j$ and $\lambda_4 = -2j$. If the sampling period is $T = 2\pi/3$ we find that

$$\mathcal{Z}[G] = \frac{3\sqrt{3}\,z}{4(z^2+z+1)}$$

and if $T = 2\pi$

$$\mathcal{Z}[G] = 0$$

Hence we see that with these sampling periods the system is not completely state controllable. For complete state controllability, $T \neq 2r\pi/3$, $T \neq r\pi/2$ ($r = \pm 1, \pm 2, \ldots$)

Notice that we can always avoid the loss of controllability by choosing the sampling period sufficiently small compared with the smallest time constant of the system.

Problem A-7-12 Consider the control system shown in Fig. 7–5. The controller is a nonlinear controller. Assuming that there is no input signal and the system is subjected to only the initial condition, determine the region of controllability for the following two cases:

Case 1: The control signal can assume only two values, $u = \pm 1$.
Case 2: The control signal can assume any value between 1 and -1, or $-1 \leqslant u \leqslant 1$.

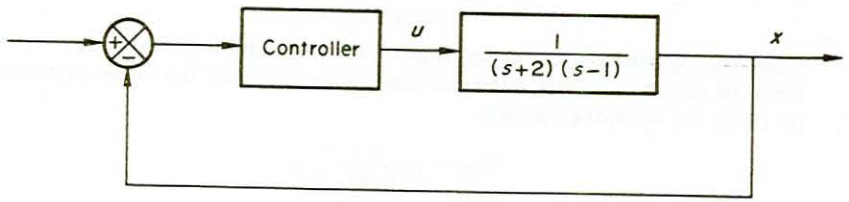


Figure 7-5 Control system.

Solution The system equation is given by

$$\ddot{x} + \dot{x} - 2x = u \qquad (7\text{–}69)$$

Case 1: For $u = 1$ this equation becomes

$$\ddot{x} + \dot{x} - 2x = 1$$

from which the equilibrium point is

$$x = -\tfrac{1}{2}, \qquad \dot{x} = 0$$

Since the characteristic equation

$$\lambda^2 + \lambda - 2 = 0$$

has roots $\lambda = 1$ and $\lambda = -2$, the equilibrium point $(-\frac{1}{2}, 0)$ is a saddle point. The separatrices which pass through the saddle point have slopes of -2 and 1. Figure 7–6(A) is a state plane diagram showing the saddle point, separatrices, and several trajectories.

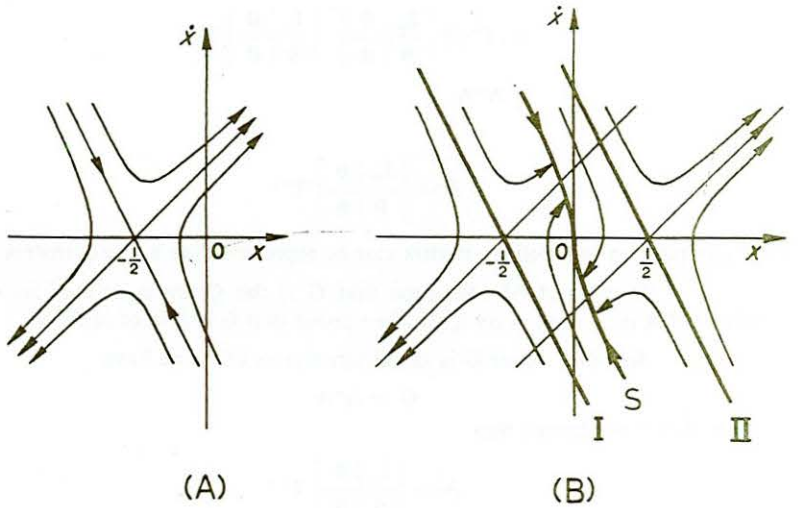


Figure 7-6 State plane diagrams for the system of Fig. 7-5. (A) Diagram showing saddle point, separatrices, and several trajectories; (B) diagram showing domain of controllability.

For $u = -1$, Eq. (7–69) becomes

$$\ddot{x} + \dot{x} - 2x = -1$$

The equilibrium point is now

$$x = \tfrac{1}{2}, \qquad \dot{x} = 0$$

It is also a saddle point. The state plane diagram is identical to the foregoing, except that it is shifted to the right by 1. Figure 7–6(B) is the superposition of the two sets of diagrams corresponding to $u = 1$ and $u = -1$.

Graphically we see that the origin can be reached from any point between separatrices I and II shown in Fig. 7–6(B). To reach the origin with $u = \pm 1$ it is necessary that the trajectory starting from a given initial point intersects the curve S. (The curve S is the switching curve consisting of the trajectory with $u = -1$ in the upper half plane and that with $u = 1$ in the lower half plane both passing through the origin.) To reach the curve S from points lying between separatrices I and II but to the left of S it is necessary to use $u = 1$ until the curve S is reached.

Similarly, for the points lying between separatrices I and II but to the right of the curve S, it is necessary to use $u = -1$ until the curve S is reached.

Graphically, it is easily seen that it is not possible to enter the region bounded by separatrices I and II from the exterior. Hence the region bounded by two separatrices I and II is the domain of controllability for $u = \pm 1$.

Case 2: The domain of controllability in this case can be determined similarly to Case 1. For u = constant and bounded by ± 1 the saddle points are

$$x = -\frac{1}{2}u, \qquad \dot{x} = 0 \qquad (-1 \leqslant u \leqslant 1)$$

For $-1 \leqslant u \leqslant 1$ (u is bounded but not necessarily constant), the maximum width of the domain of controllability is obtained when $|u|$ is maximum. Hence for Case 2 the domain of controllability is the same as Case 1 and is bounded by separatrices I and II shown in Fig. 7–6(B).

Problem A-7-13 Show that the matrix equation $\mathbf{AX} = \mathbf{B}$, where $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{X}$ are $n \times n$ constant matrices, has a unique solution if and only if

$$\text{rank } \mathbf{A} = \text{rank } [\mathbf{A} \mid \mathbf{B}] = n$$

Solution The given matrix equation is clearly equivalent to the following equation:

$$[\mathbf{AX}_1 \mid \mathbf{AX}_2 \mid \cdots \mid \mathbf{AX}_n] = [\mathbf{B}_1 \mid \mathbf{B}_2 \mid \cdots \mid \mathbf{B}_n]$$

where

$$[\mathbf{X}_1 \mid \mathbf{X}_2 \mid \cdots \mid \mathbf{X}_n] = \mathbf{X}$$

$$[\mathbf{B}_1 \mid \mathbf{B}_2 \mid \cdots \mid \mathbf{B}_n] = \mathbf{B}$$

or

$$\begin{aligned} \mathbf{AX}_1 &= \mathbf{B}_1 \\ \mathbf{AX}_2 &= \mathbf{B}_2 \\ &\cdots \\ \mathbf{AX}_n &= \mathbf{B}_n \end{aligned}$$

By Theorem 7–12 these n equations, respectively, have unique solutions if and only if

$$\begin{aligned} \text{rank } \mathbf{A} &= \text{rank } [\mathbf{A} \mid \mathbf{B}_1] = n \\ \text{rank } \mathbf{A} &= \text{rank } [\mathbf{A} \mid \mathbf{B}_2] = n \\ &\cdots \\ \text{rank } \mathbf{A} &= \text{rank } [\mathbf{A} \mid \mathbf{B}_n] = n \end{aligned}$$

Clearly $\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_n$ are linearly dependent on n columns of $\mathbf{A}$. Hence the foregoing condition is the same as

$$\text{rank } \mathbf{A} = \text{rank } [\mathbf{A} \mid \mathbf{B}] = n$$

Thus we have shown that the original matrix equation has a unique solution if and only if

$$\text{rank } \mathbf{A} = \text{rank } [\mathbf{A} \mid \mathbf{B}] = n$$

Problem A-7-14 Consider the following system:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} -1 & -2 & -2 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} [u]$$

$$y = [1 \quad 1 \quad 0] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Show that the system is completely state controllable and completely observable.

Solution For complete state controllability, the vectors $\mathbf{B}$, $\mathbf{AB}$, and $\mathbf{A}^2\mathbf{B}$ must be linearly independent, where

$$\mathbf{B} = \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}, \qquad \mathbf{AB} = \begin{bmatrix} -4 \\ 1 \\ 1 \end{bmatrix}, \qquad \mathbf{A}^2\mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ -5 \end{bmatrix}$$

Since

$$\begin{vmatrix} 2 & -4 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & -5 \end{vmatrix} = -10 \neq 0$$

it can be seen that the vectors $\mathbf{B}$, $\mathbf{AB}$, and $\mathbf{A}^2\mathbf{B}$ are linearly independent. Thus, the system is completely state controllable.

For complete observability the vectors $\mathbf{C}^*$, $\mathbf{A}^*\mathbf{C}^*$, and $(\mathbf{A}^*)^2\mathbf{C}^*$ must be linearly independent, where

$$\mathbf{C}^* = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \qquad \mathbf{A}^*\mathbf{C}^* = \begin{bmatrix} -1 \\ -3 \\ -1 \end{bmatrix}, \qquad (\mathbf{A}^*)^2\mathbf{C}^* = \begin{bmatrix} 0 \\ 5 \\ 0 \end{bmatrix}$$

Since

$$\begin{vmatrix} 1 & -1 & 0 \\ 1 & -3 & 5 \\ 0 & -1 & 0 \end{vmatrix} = 5 \neq 0$$

The vectors $\mathbf{C}^*$, $\mathbf{A}^*\mathbf{C}^*$, and $(\mathbf{A}^*)^2\mathbf{C}^*$ are linearly independent. Thus the system is completely observable.

Problem A-7-15 Consider a system defined by

$$\mathbf{x}((k+1)T) = \mathbf{G}[\mathbf{x}(kT) + \mathbf{C}^* u(kT)]$$

$$y(kT) = \mathbf{C}\mathbf{x}(kT)$$

where

$$\mathbf{G} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \qquad \mathbf{C} = [1 \quad 0 \quad 0 \quad 0]$$

Given any unknown initial state $\mathbf{x}(0)$, every state vector is reduced to zero at most 4 sampling periods if and only if the closed loop system is governed by

$$\mathbf{x}((k+1)T) = \mathbf{G}[\mathbf{x}(kT) - \mathbf{C}^*(\mathbf{C}^*, \mathbf{x}(kT))]$$

Solution The given system is completely state controllable, since column vectors $\mathbf{C}^*$, $\mathbf{GC}^*$, $\mathbf{G}^2\mathbf{C}^*$, and $\mathbf{G}^3\mathbf{C}^*$ are linearly independent. By Theorem 7–5, the fastest control scheme is uniquely determined by

$$u(kt) = -(\eta_1, \mathbf{x}(kT))$$

where η_1 is the first vector in the dual basis $\eta_1, \eta_2, \ldots, \eta_n$ of the basis $\xi_1, \xi_2, \ldots, \xi_n$ where $\xi_i = \mathbf{G}^{-i+1}\mathbf{C}^*$. For this system η_1 is found to be $\mathbf{C}^*$. Hence

$$u(kT) = -(\mathbf{C}^*, \mathbf{x}(kT))$$

which is the negative of the output $y(kT)$. The following scheme:

$$\mathbf{x}((k+1)T) = \mathbf{G}[\mathbf{x}(kT) - \mathbf{C}^*(\mathbf{C}^*, \mathbf{x}(kT))]$$

will reduce every state vector to zero at most 4 sampling periods.

Problem A-7-16 Consider the system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 3 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

The output is given by

$$y = [1 \quad 1 \quad 1] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

(1) Show that the system is not completely observable.
(2) Show also that if the output is given by

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

the system is completely observable.

Solution

(1) Let us define

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 3 & 1 \end{bmatrix}, \qquad \mathbf{C} = [1 \quad 1 \quad 1]$$

Then

$$[\mathbf{C}^* \,|\, \mathbf{A}^*\mathbf{C}^* \,|\, (\mathbf{A}^*)^2\mathbf{C}^*] = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 5 & 13 \\ 1 & 1 & 1 \end{bmatrix}$$

Since the rank of this matrix is two, the system is not completely observable. (Notice that the degree of the minimal polynomial of $\mathbf{A}$ is two. Hence $(\mathbf{A}^*)^2\mathbf{C}^*$ is linearly dependent on $\mathbf{C}^*$ and $\mathbf{A}^*\mathbf{C}^*$.)

(2) Let us define

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}$$

Then

$$[\mathbf{C}^* \,|\, \mathbf{A}^*\mathbf{C}^* \,|\, (\mathbf{A}^*)^2\mathbf{C}^*] = \left[\begin{array}{cc|cc|cc} 1 & 1 & 2 & 2 & 4 & 4 \\ 1 & 2 & 5 & 13 & 13 & 35 \\ 1 & 3 & 1 & 3 & 1 & 3 \end{array}\right]$$

Clearly

$$\begin{vmatrix} 1 & 1 & 2 \\ 1 & 2 & 5 \\ 1 & 3 & 1 \end{vmatrix} \neq 0$$

Hence the rank of this 3×6 matrix is three. The system is, therefore, completely observable.

Problem A-7-17 (7–2) Consider the linear time-varying system

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u} \tag{7–70}$$

where
$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{A}(t) = n \times n$ matrix, elements of which are piecewise continuous functions of time
$\mathbf{B}(t) = n \times r$ matrix, elements of which are piecewise continuous functions of time

Prove that the system is completely state controllable on $t_0 \leqslant t \leqslant t_1$ if and only if

$$\mathbf{B}^*(t)\Phi^*(t_1, t)\boldsymbol{\lambda} \neq \mathbf{0}$$

on $t_0 \leqslant t \leqslant t_1$ for each $\boldsymbol{\lambda} \neq \mathbf{0}$ in the state space V, where $\Phi(t, t_0)$ is the unique fundamental matrix satisfying

$$\frac{d}{dt}\Phi(t, t_0) = \mathbf{A}(t)\Phi(t, t_0), \qquad \Phi(t_0, t_0) = \mathbf{I}$$

and $\boldsymbol{\lambda}$ is a fixed length nonzero vector (pointing in every direction in the state space V).

Solution If the system is completely state controllable then any initial state can be transferred to any final state in a finite time interval $t_0 \leqslant t \leqslant t_1$. We shall first convert the problem of transferring to $\mathbf{x}(t_1)$ from some initial state $\mathbf{x}(t_0)$ to that of transferring to $\mathbf{x}^d(t_1)$ from the origin of the state space V. The solution of Eq. (7–70) is given by

$$\mathbf{x}(t) = \Phi(t, t_0)\mathbf{x}(t_0) + \int_{t_0}^{t} \Phi(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau$$

Define

$$\mathbf{x}^d(t) = \mathbf{x}(t) - \Phi(t, t_0)\mathbf{x}(t_0)$$

Then,

$$\mathbf{x}^d(t) = \int_{t_0}^{t} \Phi(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)d\tau \tag{7-71}$$

The problem of transferring to $\mathbf{x}(t_1)$ from some initial state $\mathbf{x}(t_o)$ is thus converted to that of transferring to $\mathbf{x}^d(t_1)$ from the origin of the state space V.

Notice that the system is completely state controllable on $t_0 \leqslant t \leqslant t_1$ if and only if every point $\mathbf{x}^d(t_1)$ in the state space can be reached from the origin, or in other words, if the region $\Gamma(t_0, t_1)$ of all the points $\mathbf{x}^d(t_1)$ in Eq. (7–71) which can be reached from the origin coincides with the whole state space V. Notice that $\Gamma(t_0, t_1)$ is convex. Hence, in order to show that $\Gamma(t_0, t_1)$ coincides with the whole state space V, we need show only that $\Gamma(t_0, t_1)$ is not bounded in any direction in the state space. Let us take the inner product of $\mathbf{x}^d(t_1)$ with every fixed-length nonzero vector $\boldsymbol{\lambda}$ (pointing in every direction in the state space V) and show that it can be made as large as desired by some choice of $\mathbf{u}$.

$$(\boldsymbol{\lambda}, \mathbf{x}^d(t_1)) = \int_{t_0}^{t_1} (\boldsymbol{\lambda}, \Phi(t_1, t)\mathbf{B}(t)\mathbf{u}(t))\, dt \tag{7-72}$$

If $\mathbf{u}(t)$ is chosen such that

$$\mathbf{u}(t) = k\mathbf{B}^*(t)\Phi^*(t_1, t)\boldsymbol{\lambda}$$

where k is positive, then Eq. (7–72) becomes

$$(\boldsymbol{\lambda}, \mathbf{x}^d(t_1)) = k \int_{t_0}^{t_1} (\mathbf{B}^*(t)\Phi^*(t_1, t)\boldsymbol{\lambda}, \mathbf{B}^*(t)\Phi^*(t_1, t)\boldsymbol{\lambda})\, dt \tag{7-73}$$

A sufficient condition for complete state controllability on $t_0 \leqslant t \leqslant t_1$ is that

$$\mathbf{B}^*(t)\Phi^*(t_1, t)\boldsymbol{\lambda} \neq \mathbf{0} \tag{7-74}$$

on $t_0 \leqslant t \leqslant t_1$ for each $\boldsymbol{\lambda} \neq \mathbf{0}$ in the state space V, because then the integral in Eq. (7–73) is positive for every $\boldsymbol{\lambda} \neq \mathbf{0}$, and $(\boldsymbol{\lambda}, \mathbf{x}^d(t_1))$ can be made as large as desired by increasing k. The condition (7–74) is also necessary because if it is violated by some $\boldsymbol{\lambda} = \boldsymbol{\lambda}_1$, then no matter how $\mathbf{u}(t)$ is chosen in Eq. (7–72), $(\boldsymbol{\lambda}_1, \mathbf{x}^d(t_1)) = 0$ and no point $\mathbf{x}^d(t_1)$ in the direction of $\boldsymbol{\lambda}_1$ can be reached. This completes the proof.

Problem A-7-18 Consider the system defined by

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$$

where
$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{A} = n \times n$ constant matrix
$\mathbf{B} = n \times r$ constant matrix

A necessary and sufficient condition for complete state controllability of this system can be stated thus: the Gram matrix given by Eq. (7–47) or

$$\mathbf{W}(0, t_1) = \int_0^{t_1} \Phi(0, t)\mathbf{B}\mathbf{B}^*\Phi^*(0, t)\, dt$$

where

$$\Phi(0, t) = e^{-\mathbf{A}t}$$

is nonsingular. Prove that this condition is equivalent to the statement that the composite $n \times nr$ matrix $\mathbf{P}$ where

$$\mathbf{P} = [\mathbf{B} \mid \mathbf{A}\mathbf{B} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{B}]$$

is of rank n.

Solution The Gram matrix

$$\mathbf{W}(0, t_1) = \int_0^{t_1} \Phi(0, t)\mathbf{B}\mathbf{B}^*\Phi^*(0, t)\, dt$$

is nonsingular implies that the n columns of $\mathbf{B}^*\Phi^*(0, t)$ are linearly independent, or n rows of $\Phi(0, t)\mathbf{B}$ are linearly independent. Notice that in the present problem $\mathbf{A}$ is a constant matrix. Hence

$$\Phi(0, t)\mathbf{B} = e^{-\mathbf{A}t}\mathbf{B} = \sum_{i=0}^{p-1} \alpha_i(t)\mathbf{A}^i\mathbf{B}$$

where p is the degree of the minimal polynomial of $\mathbf{A}$ and $\alpha_0(t), \alpha_1(t), \ldots, \alpha_{p-1}(t)$ are linearly independent. Since

$$\Phi(0, t)\mathbf{B} = [\mathbf{B} \mid \mathbf{A}\mathbf{B} \mid \cdots \mid \mathbf{A}^{p-1}\mathbf{B}] \begin{bmatrix} \alpha_0(t) \\ \alpha_1(t) \\ \vdots \\ \alpha_{p-1}(t) \end{bmatrix}$$

linear independence of the n rows of $\Phi(0, t)\mathbf{B}$ implies that the rank of

$$[\mathbf{B} \mid \mathbf{A}\mathbf{B} \mid \cdots \mid \mathbf{A}^{p-1}\mathbf{B}], \qquad p \leqslant n$$

is n which is equivalent to the statement that the rank of

$$[\mathbf{B} \mid \mathbf{A}\mathbf{B} \mid \cdots \mid \mathbf{A}^{n-1}\mathbf{B}]$$

is n.

PROBLEMS

Problem B-7-1 Suppose that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are linearly independent column vectors and $\mathbf{A}$ is a nonsingular matrix. Show that $\mathbf{A}\mathbf{x}_1, \mathbf{A}\mathbf{x}_2, \ldots, \mathbf{A}\mathbf{x}_n$ are also linearly independent.

Problem B-7-2 Consider a linear time-varying system

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u}$$

where
$\mathbf{x} = n$ vector
$\mathbf{u} = r$ vector
$\mathbf{A}(t) = n \times n$ matrix, elements of which are piecewise continuous functions of time
$\mathbf{B}(t) = n \times r$ matrix, elements of which are piecewise continuous functions of time

Show that the solution of this equation can be given by

$$\mathbf{x}(t) = \mathbf{Y}^{-1}(t, t_0)\left[\mathbf{x}(t_0) + \int_{t_0}^{t} \mathbf{Y}(\tau, t_0)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau\right]$$

where $\mathbf{Y}(t, t_0)$ is the solution of

$$\frac{d}{dt}\mathbf{Y}(t, t_0) = -\mathbf{Y}(t, t_0)\mathbf{A}(t), \qquad \mathbf{Y}(t_0, t_0) = \mathbf{I}$$

Problem B-7-3 Consider the system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ -1 \end{bmatrix}[u]$$

$$y = [1 \quad 1]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Show that the output y is independent of the control function u.

Problem B-7-4 Suppose that a system described by

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u$$

where
$\mathbf{x} = n$ vector
$u =$ scalar
$\mathbf{A} = n \times n$ matrix
$\mathbf{B} = n \times 1$ matrix

is completely state controllable. Prove that the system described by the following equation is also completely state controllable.

$$\dot{\mathbf{y}} = \mathbf{K}^{-1}\mathbf{A}\mathbf{K}\mathbf{y} + \mathbf{K}^{-1}\mathbf{B}u$$

where $\mathbf{K}$ is an arbitrary $n \times n$ constant matrix which is nonsingular.

Problem B-7-5 Consider a system described by the following scalar differential equation:

$$\overset{(n)}{x} + a_1\overset{(n-1)}{x} + \cdots + a_n x = b_0\overset{(m)}{u} + b_1\overset{(m-1)}{u} + \cdots + b_m u, \qquad (m \leqslant n)$$

Find the condition that for any nonzero initial conditions $x(0)$, $\dot{x}(0), \ldots, \overset{(n-1)}{x}(0)$, we have $x(t \geqslant t_1) = 0$ for some $t_1 > 0$.

Problem B-7-6 Consider the system described by

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u}$$
$$\mathbf{y} = \mathbf{C}(t)\mathbf{x} + \mathbf{D}(t)\mathbf{u}$$

where
$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{y} = m$ vector (output vector) $(m \leqslant n)$
$\mathbf{A}(t) = n \times n$ matrix
$\mathbf{B}(t) = n \times r$ matrix
$\mathbf{C}(t) = m \times n$ matrix
$\mathbf{D}(t) = m \times r$ matrix

and the elements of $\mathbf{A}(t)$, $\mathbf{B}(t)$, $\mathbf{C}(t)$, and $\mathbf{D}(t)$ are piecewise continuous functions of time. The output can be given by

$$\mathbf{y}(t) = \mathbf{C}(t)\Phi(t, t_0)\mathbf{x}(t_0) + \int_{t_0}^{t} \mathbf{H}_y(t, \tau)\mathbf{u}(\tau)\, d\tau + \mathbf{D}(t)\mathbf{u}(t)$$

where $\mathbf{H}_y(t, \tau) = \mathbf{C}(t)\Phi(t, \tau)\mathbf{B}(\tau)$

and $\Phi(t, t_0)$ satisfies the equation

$$\frac{d}{dt}\Phi(t, t_0) = \mathbf{A}(t)\Phi(t, t_0), \qquad \Phi(t_0, t_0) = \mathbf{I}$$

Find the condition that the system is completely output controllable on $t_0 \leqslant t \leqslant t_1$. Write the condition in terms of an appropriate Gram matrix involving $\mathbf{H}_y(t, \tau)$ and $\mathbf{D}(t)$.

Problem B-7-7 (7–8) Consider the system

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)u \tag{7–75}$$

where
$\mathbf{x} = n$ vector (state vector)
$u =$ scalar (control signal)
$\mathbf{A}(t) = n \times n$ matrix, elements of which are piecewise continuous functions of time
$\mathbf{B}(t) = n \times 1$ matrix, elements of which are piecewise continuous functions of time

The solution of Eq. (7–75) is

$$\mathbf{x}(t) = \Phi(t, t_0)\mathbf{x}(t_0) + \int_{t_0}^{t} \Phi(t, \tau)\mathbf{B}(\tau)u(\tau)\, d\tau$$
$$= \Phi(t, t_0)\mathbf{x}(t_0) + \int_{t_0}^{t} \mathbf{h}(t, \tau)u(\tau)\, d\tau$$

where $\mathbf{h}(t, \tau)$ is the response vector of the system to an impulse $u(t) = \delta(t - \tau)$ applied at $t = \tau$. Assume that the functions $h_i(t, \tau)$ are linearly independent on all time intervals, $t_0 \leqslant \tau \leqslant t$. Namely, for every nonzero n vector $\boldsymbol{\lambda}$ the inner product of $\boldsymbol{\lambda}$ and $\mathbf{h}(t, \tau)$ is nonzero, or

$$(\boldsymbol{\lambda}, \mathbf{h}(t, \tau)) \neq 0 \qquad \text{on} \qquad t_0 \leqslant \tau \leqslant t$$

Prove that this condition is equivalent to the statement that the system can be brought by a suitable control signal u from any initial state at $t = t_0$ to any final state $\mathbf{x}(t_1)$ at $t = t_1$ on an arbitrary finite time interval, $t_0 \leqslant t \leqslant t_1$.

Problem B-7-8 Consider the system defined by

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$$
$$\mathbf{y} = \mathbf{C}\mathbf{x}$$

where
$\mathbf{x} = n$ vector (state vector)
$\mathbf{y} = m$ vector (output vector)
$\mathbf{A} = n \times n$ constant matrix
$\mathbf{C} = m \times n$ constant matrix

A necessary and sufficient condition for complete observability of this system can be stated that the following Gram matrix be nonsingular:

$$\mathbf{M}(0, t_1) = \int_0^{t_1} \Phi^*(t, 0)\mathbf{C}^*\mathbf{C}\Phi(t, 0)\, dt$$

where $\Phi(t, 0) = e^{\mathbf{A}t}$

Prove that this condition is equivalent to the statement that the composite $n \times mn$ matrix $\mathbf{P}$ where

$$\mathbf{P} = [\mathbf{C}^* \,|\, \mathbf{A}^*\mathbf{C}^* \,|\, \cdots \,|\, (\mathbf{A}^*)^{n-1}\mathbf{C}^*]$$

is of rank n.

Problem B-7-9 Consider the system

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u} \\ \mathbf{y} &= \mathbf{C}(t)\mathbf{x} \end{aligned} \tag{7-76}$$

Prove that if $\mathbf{z}(t) = \mathbf{0}$ (for $t \geqslant t_0$) is the only solution to the following two simultaneous equations:

$$\dot{\mathbf{z}} = -\mathbf{A}^*(t)\mathbf{z}, \qquad \mathbf{B}^*(t)\mathbf{z} = \mathbf{0}$$

then the system defined by Eq. (7–76) is completely state controllable.

Prove also that if $\mathbf{x}(t) = \mathbf{0}$ (for $t \geqslant t_0$) is the only solution to the following two simultaneous equations:

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}, \qquad \mathbf{C}(t)\mathbf{x} = \mathbf{0}$$

then the system defined by Eq. (7–76) is completely observable.

REFERENCES

7–1 Kalman, R. E., "On the General Theory of Control Systems," *Proc. First Intern. Cong.* IFAC, Moscow, 1960, *Automatic and Remote Control*, London: Butterworths & Company Ltd., 1961, pp. 481–92.

7–2 Kreindler, E., and P. E. Sarachik, "On the Concepts of Controllability and Observability of Linear Systems," *IEEE Trans. Automatic Control* (1964), 129–36.

7–3 Kalman, R. E., Y. C. Ho, and K. S. Narendra, "Controllability of Linear Dynamical Systems," *Cont. Differential Equations*, **1** (1962), 189–213.

7–4 Falb, P. L., and M. Athans, "A Direct Constructive Proof of the Criterion for Complete Controllability of Time-Invariant Linear Systems," *IEEE Trans. Automatic Control* (1964), 189–90.

7–5 Butman, S., and R. Sivan (Sussman), "On Cancellations, Controllability and Observability," *IEEE Trans. Automatic Control* (1964), 317–18.

7–6 Kalman, R. E., "Canonical Structure of Linear Dynamical Systems," *Proc. Natl. Acad. Sci. USA* **48** (1962), 596–600.

7–7 Zadeh, L. A., and C. A. Desoer, *Linear System Theory—The State Space Approach*, New York: McGraw-Hill Book Company, Inc., 1963, pp. 505–09.

7–8 Kreindler, E., "Contributions to the Theory of Time Optimal Control," *J. Franklin Inst.*, **275** (1963), 314–44.

8

Stability Analysis

8-1 INTRODUCTION (8–1, 8–2, 8–3)[1]

The degree of complexity of stability analysis of dynamic systems increases rapidly as the description of system differential equations is modified from linear time-invariant equations to linear time-varying equations and to nonlinear equations. A variety of stablility criteria is applicable to linear time-invariant systems. Only several methods are applicable to stability analysis of linear time-varying systems and nonlinear systems. Most of these are quite limited in their applications.

This chapter presents Routh-Hurwitz stability criteria which are applicable to linear time-invariant systems and the Liapunov methods of stability analysis (the first method of Liapunov and the second method of Liapunov) which are applicable to both linear and nonlinear systems. We shall devote much attention to the second method of Liapunov, which provides stability information on linear and nonlinear differential equations without solving them; hence Liapunov's second method is called the *direct method of Liapunov*. Because, in general, explicit solutions of nonlinear differential equations are wholly out of the question, the second method of Liapunov is most useful for investigating the stability of nonlinear systems. This method gives sufficient conditions for asymptotic stability of equilibrium states of nonlinear systems and gives necessary and sufficient conditions for asymptotic stability of equilibrium states of linear time-invariant systems.

[1]Figures in parentheses refer to material listed at the end of this chapter.

In addition to stability analysis the second method of Liapunov is quite useful in designing optimal control systems that are stable.

In the following portions of this section we give definitions of several terminologies necessary for stability analysis of dynamic systems.

System The system we consider in this chapter is defined by

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t) \tag{8-1}$$

where $\mathbf{x}$ is a state vector (n vector) and $\mathbf{f}(\mathbf{x}, t)$ is an n vector whose elements are functions of $x_1, x_2, \ldots, x_n$ and t. We shall assume uniqueness of solutions and a continuous dependence of solutions on initial conditions. We shall denote the solution of Eq. (8–1) as $\phi(t; \mathbf{x}_0, t_0)$, where $\mathbf{x} = \mathbf{x}_0$ at $t = t_0$ and t is the observed time. Thus,

$$\phi(t_0; \mathbf{x}_0, t_0) = \mathbf{x}_0$$

Motion In defining stability, asymptotic stability, and instability we shall frequently use the word "motion." A *motion* is defined as a trajectory starting from any state or any point in the n-dimensional state space.

Equilibrium states In the system of Eq. (8–1) a state $\mathbf{x}_e$ where

$$\mathbf{f}(\mathbf{x}_e, t) = \mathbf{0} \qquad \text{for all } t \tag{8-2}$$

is called an *equilibrium state* of the system. If the system is linear time-invariant, namely $\mathbf{f}(\mathbf{x}, t) = \mathbf{A}\mathbf{x}$, then there exists only one equilibrium state if $\mathbf{A}$ is nonsingular, and there exist infinitely many equilibrium states if $\mathbf{A}$ is singular. For nonlinear systems there may be one or more than one equilibrium state. The equilibrium states correspond to the constant solutions of the system. ($\mathbf{x} = \mathbf{x}_e$ for all t.) The determination of the equilibrium states does not involve the solution of the system's differential equation, Eq. (8–1), but only the solution of Eq. (8–2).

If equilibrium states are isolated from each other, they are called *isolated* equilibrium states. Any isolated equilibrium state can be shifted to the origin of the coordinates, or $\mathbf{f}(\mathbf{0}, t) = \mathbf{0}$, by a translation of coordinates. In this book we shall treat stability analysis of only isolated equilibrium states. Therefore, in this book an equilibrium state $\mathbf{x}_e$ should be interpreted as an isolated equilibrium state. (Most of the existing stability theorems are concerned with stability or asymptotic stability or instability of isolated equilibrium states.)

Stability An equilibrium state $\mathbf{x}_e$ of the system of Eq. (8–1) is said to be stable if for each real number $\epsilon > 0$ there is a real number $\delta(\epsilon, t_0) > 0$ such that the inequality

$$\|\mathbf{x}_0 - \mathbf{x}_e\| \leqslant \delta$$

implies

$$\|\phi(t; \mathbf{x}_0, t_0) - \mathbf{x}_e\| \leqslant \epsilon \qquad \text{for all } t \geqslant t_0$$

The real number δ depends on ϵ and, in general, also depends on t_0. If δ does not depend on t_0, the equilibrium state is said to be *uniformly stable*. Figure 8–1 shows a stable equilibrium state $\mathbf{x}_e$ of a second order system and the representative trajectory starting from $\mathbf{x}_0$. In Fig. 8–1 $S(\epsilon)$ and $S(\delta)$ are

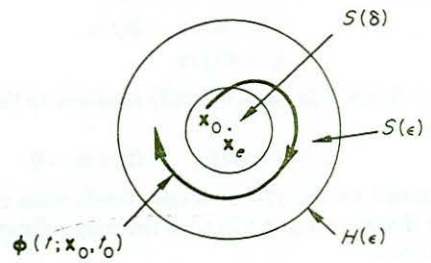


Figure 8-1 Stable equilibrium state $\mathbf{x}_e$ and representative trajectory starting from $\mathbf{x}_0$.

respectively circular regions of radii $\epsilon > 0$ and $\delta > 0$ about the equilibrium state $\mathbf{x}_e$. [In the case of an n-dimensional state space any closed, bounded surfaces can be used in place of circles to define regions $S(\epsilon)$ and $S(\delta)$.] $S(\epsilon)$ consists of the states $\mathbf{x}$ satisfying $\|\mathbf{x} - \mathbf{x}_e\| < \epsilon$. This figure shows that corresponding to each $S(\epsilon)$ there is an $S(\delta)$ such that a motion starting at a state $\mathbf{x}_0$ in $S(\delta)$ does not leave $S(\epsilon)$, or $\mathbf{x}_0$ in $S(\delta)$ implies $\phi(t; \mathbf{x}_0, t_0)$ is in $S(\epsilon)$ for all $t \geqslant t_0$. That is, a motion starting in $S(\delta)$ never reaches the boundary circle $H(\epsilon)$ of $S(\epsilon)$.

The solutions of Eq. (8–1) are said to be *bounded* if there exists for a given $\delta > 0$ a constant $\epsilon(\delta, t_0)$ such that

$$\|\mathbf{x}_0 - \mathbf{x}_e\| \leqslant \delta$$

implies

$$\|\phi(t; \mathbf{x}_0, t_0) - \mathbf{x}_e\| \leqslant \epsilon(\delta, t_0) \qquad \text{for all } t \geqslant t_0$$

If ϵ does not depend on t_0, the solutions are said to be uniformly bounded.

Asymptotic stability An equilibrium state $\mathbf{x}_e$ of the system given by Eq. (8–1) is said to be *asymptotically stable* if it is stable and if every solution starting at a state $\mathbf{x}_0$ sufficiently near $\mathbf{x}_e$ converges to $\mathbf{x}_e$ as t increases indefinitely. Namely, given two real numbers $\epsilon > 0$ and $\mu > 0$ there are real numbers $\delta > 0$ and $T(\mu, \delta, t_0)$ such that

$$\|\mathbf{x}_0 - \mathbf{x}_e\| \leqslant \delta$$

implies

$$\|\phi(t; \mathbf{x}_0, t_0) - \mathbf{x}_e\| \leqslant \epsilon \qquad \text{for all } t \geqslant t_0$$

and

$$\|\phi(t; \mathbf{x}_0, t_0) - \mathbf{x}_e\| \leqslant \mu \qquad \text{for all } t \geqslant t_0 + T(\mu, \delta, t_0)$$

Figure 8–2(A) shows an asymptotically stable equilibrium state $\mathbf{x}_e$ of a second order system and the representative trajectory starting from $\mathbf{x}_0$. Figure 8–2(B) is a plot of the representative trajectory as time elapses. In these figures, $S(\epsilon)$, $S(\delta)$, and $S(\mu)$ are, respectively, circular regions of radii $\epsilon > 0$, $\delta > 0$, and $\mu > 0$ about the equilibrium state $\mathbf{x}_e$. If the equilibrium state $\mathbf{x}_e$ is asymptotically stable, then every motion starting at a state $\mathbf{x}_0$ in $S(\delta)$ converges, without leaving $S(\epsilon)$, to the origin as time increases indefinitely.

In practice, asymptotic stability is more important than mere stability. Since asymptotic stability is a local concept, merely to have established asymptotic stability may not mean that the system will operate properly. Some knowledge of the size of the largest region of asymptotic stability is usually necessary. Such largest region of asymptotic stability is called the

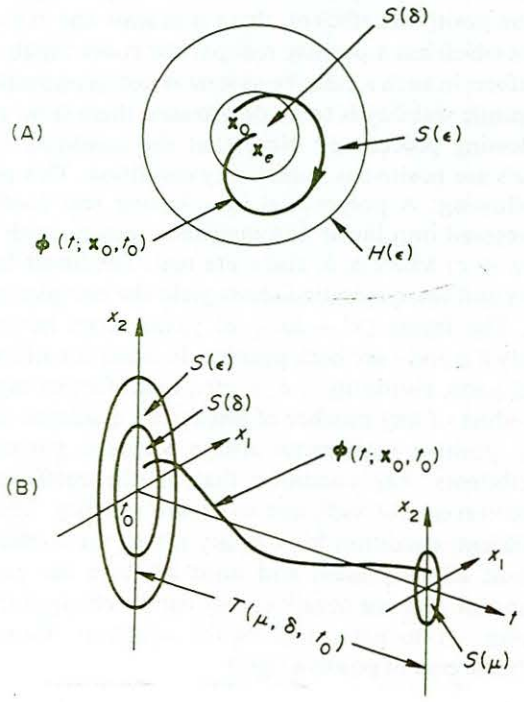


Figure 8-2 (A) Asymptotically stable equilibrium state $\mathbf{x}_e$ and representative trajectory starting from $\mathbf{x}_0$; (B) plot of representative trajectory as time elapses.

domain of attraction. It is that part of the state space in which asymptotically stable motions originate. In other words, every motion originating in the domain of attraction is asymptotically stable.

Asymptotic stability in the large If asymptotic stability holds for all states (all points in the state space) from which motions originate, the equilibrium state is said to be *asymptotically stable in the large*. Namely, the equilibrium state $\mathbf{x}_e$ of the system given by Eq. (8–1) is said to be asymptotically stable in the large if it is stable and if every solution converges to $\mathbf{x}_e$ as t increases indefinitely. Obviously a necessary condition for asymptotic stability in the large is that there be only one equilibrium state in the whole state space.

In control engineering problems, the desirable feature is asymptotic stability in the large. If the equilibrium state of a system is not asymptotically stable in the large, then the problem becomes that of determining the largest region of asymptotic stability. This is usually very difficult. For practical purposes, however, it is sufficient to determine a region of asymptotic stability large enough so that no disturbance will exceed it.

Instability An equilibrium state is said to be *unstable* if it is neither stable nor asymptotically stable. Figure 8–3 shows an unstable equilibrium state $\mathbf{x}_e$ of a second order system and a representative trajectory starting from $\mathbf{x}_0$. As may be seen from Fig. 8–3, in the case of an unstable equilibrium state, for some real number $\epsilon > 0$ and any real number $\delta > 0$, no matter how small, there is always in the circular region $S(\delta)$ a state $\mathbf{x}_0$ such that the motion starting from this state reaches the boundary circle $H(\epsilon)$ of $S(\epsilon)$.

The foregoing definitions are the minimum requirement for understanding the stability analysis of linear and nonlinear systems presented in this chapter. Note that these definitions are not the only ones defining concepts of stability of an equilibrium state. In fact, various other ways to define concepts

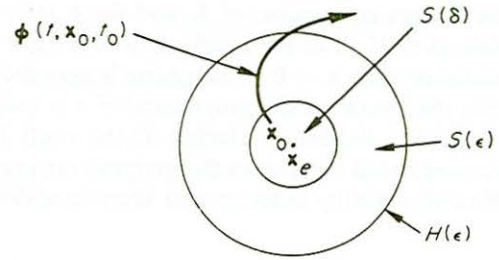


Figure 8-3 Unstable equilibrium state $\mathbf{x}_e$ and representative trajectory starting from $\mathbf{x}_0$.

of stability are available in the literature. In this chapter, however, our discussions on stability analysis will be based on the definitions just given.

Outline of the chapter The following sections of this chapter present classical Routh-Hurwitz stability criteria applicable to linear time-invariant systems (Section 8–2); the first method of Liapunov (Section 8–3); fundamental theorems of the second method of Liapunov (Section 8–4); stability analysis of linear time-invariant systems by the second method of Liapunov (Section 8–5); a proof of Routh-Hurwitz stability criteria (Section 8–5); stability analysis of nonlinear systems (Section 8–6); stability analysis of discrete-time systems (Section 8–7); a brief discussion of applications of the second method of Liapunov to system design (Section 8–8). Section 8–9 presents some concluding comments.

8-2 ROUTH'S AND HURWITZ STABILITY CRITERIA

Stability analysis of linear time-invariant systems is important not only because many practical systems can be approximated by linear time-invariant models but also because local stability of an equilibrium state of a nonlinear system may be examined by stability of a linear time-invariant model which may be obtained by use of a linearization technique.

Consider the system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$$

where $\mathbf{x}$ is a state vector (n vector) and $\mathbf{A}$ is an $n \times n$ nonsingular martix. The equilibrium state is clearly the origin, $\mathbf{x} = \mathbf{0}$. The nature of the solution of this equation is determined from the eigenvalues of $\mathbf{A}$, or the roots of the characteristic equation

$$|\mathbf{A} - \lambda\mathbf{I}| = 0$$

If the coefficients in the characteristic equation are all real, then the roots of this equation are either real or complex conjugate. If the number of the distinct roots is p, then the n components of the solution $\mathbf{x}$ can be written as

$$x_i = \sum_{j=1}^{p} g_{ij}(t)e^{\lambda_j t} \qquad (i = 1, 2, \ldots, n)$$

where the λ_j are the distinct eigenvalues of $\mathbf{A}$, and the $g_{ij}(t)$ are either constants or polynomials in t. If all of the eigenvalues of $\mathbf{A}$ have negative real parts, then the equilibrium state $\mathbf{x} = \mathbf{0}$ of the system is asymptotically stable.

Once we calculate the characteristic polynomial of $\mathbf{A}$ a variety of stability criteria can be applied to determine whether all the roots have negative real parts. Two commonly used criteria for this purpose are presented in this section. They are Routh's stability criterion and Hurwitz stability criterion. (Although these two stability criteria may seem different from each other they are actually equivalent and furnish the same information on stability.)

In the following we first present Routh's stability criterion and then Hurwitz stability criterion. A proof of these stability criteria is given in Section 8–5.

Routh's stability criterion Consider the following polynomial with constant coefficients:

$$a_0\lambda^n + a_1\lambda^{n-1} + a_2\lambda^{n-2} + \cdots + a_{n-1}\lambda + a_n = 0 \qquad (8\text{–}3)$$

where coefficients a's are real quantities. [Often Eq. (8–3) is divided by a_0, making the coefficient of the λ^n term unity.)

If a polynomial with constant but undetermined coefficients is given, Routh's stability criterion gives the condition for the coefficients of the polynomial to be satisfied in order that all the roots have negative real parts. [This is the condition for the equilibrium state of the system whose characteristic equation is given by Eq. (8–3) to be asymptotically stable.] If the polynomial with known constant coefficients is given, Routh's stability criterion determines how many roots of the polynomial have negative real parts, how many have positive real parts, and how many have zero real parts.

The procedure for stability analysis by this criterion is as follows:

1. If any coefficients a's of Eq. (8–3) are zero or negative in the presence of at least one positive coefficient, there is at least one zero root, or at least one root which has a positive real part or roots which are imaginary. (Therefore, in such a case the system is not asymptotically stable and if asymptotic stability is to be determined there is no need to follow the following procedure.) Note that the condition that all the coefficients a's are positive is a necessary condition. This may be seen from the following: A polynomial in λ having real coefficients can always be factored into linear and quadratic factors, such as $(\lambda + a)$ and $(\lambda^2 + b\lambda + c)$ where a, b, and c are real. The linear factors yield the real roots and the quadratic factors yield the complex roots of the polynomial. The factor $(\lambda^2 + b\lambda + c)$ yields roots having negative real parts only if b and c are both positive. In order for all roots to have negative real parts, constants a, b, c, etc., in all factors must be positive. The product of any number of linear and quadratic factors containing only positive coefficients always yields a polynomial with positive coefficients. The condition that all the coefficients a's are positive is, however, not sufficient to insure stability. The necessary (but not sufficient) condition for stability is that the coefficients a's of Eq. (8–3) must all be present and must all have the positive sign. (If all the constants a's are negative, they can be changed into positive by mutliplying -1 to both sides of the equation. Hence we need consider only the case of positive sign.)

2. If all the coefficients a's are positive, arrange the coefficients of the polynomial in rows and columns according to the following pattern:

$$\begin{array}{lllllll}
\lambda^n & a_0 & a_2 & a_4 & a_6 & \cdots & \leftarrow \text{first row} \\
\lambda^{n-1} & a_1 & a_3 & a_5 & a_7 & \cdots & \leftarrow \text{second row} \\
\lambda^{n-2} & b_1 & b_2 & b_3 & b_4 & \cdots & \\
\lambda^{n-3} & c_1 & c_2 & c_3 & c_4 & \cdots & \\
\vdots & \vdots & \vdots & \vdots & & & \\
\lambda^3 & d_1 & d_2 & 0 & & & \\
\lambda^2 & e_1 & e_2 & 0 & & & \\
\lambda^1 & f_1 & 0 & & & & \\
\lambda^0 & g_1 & & & & &
\end{array}$$

$\uparrow$
first column

The constants b_1, b_2, b_3, etc., are evaluated according to the following schedule:

$$b_1 = \frac{a_1 a_2 - a_0 a_3}{a_1}$$

$$b_2 = \frac{a_1 a_4 - a_0 a_5}{a_1}$$

$$b_3 = \frac{a_1 a_6 - a_0 a_7}{a_1}$$

$$\cdots$$

$$\cdots$$

The evaluation for b's is continued until the rest of the b's are all equal to zero. The same schedule of cross multiplying the coefficients of the two previous rows is used for evaluating the coefficients c's, d's, e's, etc. Namely,

$$c_1 = \frac{b_1 a_3 - a_1 b_2}{b_1}$$

$$c_2 = \frac{b_1 a_5 - a_1 b_3}{b_1}$$

$$c_3 = \frac{b_1 a_7 - a_1 b_4}{b_1}$$

$$\cdots$$

$$\cdots$$

$$f_1 = \frac{e_1 d_2 - d_1 e_2}{e_1}$$

and

$$g_1 = e_2$$

(Evaluation of the coefficients is continued until the $(n + 1)$th row is completed. The Routh array of coefficients is triangular.) It is noted that in developing the array the entire row may be divided or multiplied by a positive number in order to simplify the numerical calculation without altering the results on stability. Routh's stability criterion states that the number of roots (of the polynomial) with positive real parts is equal to the number of changes of sign of the coefficients in the first column. Note that the exact values of the terms in the first column need not be known, instead only signs of the terms are needed. The necessary and sufficient condition that all the roots of Eq. (8–3) have negative real parts is that all the coefficients a's of Eq. (8–3) are positive and all the terms in the first column of the array have positive signs.

Example 8-1 If Routh's stability criterion is applied to the following third-degree polynomial:

$$a_0\lambda^3 + a_1\lambda^2 + a_2\lambda + a_3 = 0$$

where all the coefficients a's are positive numbers, then the Routh array of coefficients becomes

$$\begin{array}{lll}
\lambda^3 & a_0 & a_2 \\
\lambda^2 & a_1 & a_3 \\
\lambda^1 & \dfrac{a_1 a_2 - a_0 a_3}{a_1} & 0 \\
\lambda^0 & a_3 &
\end{array}$$

The condition that all the roots have negative real parts is given by $a_1 a_2 > a_0 a_3$. As the degree of the polynomial increases, the conditions for asymptotic stability become more complicated.

Example 8-2 Consider the following example:

$$\lambda^4 + 2\lambda^3 + 3\lambda^2 + 4\lambda + 5 = 0$$

Following the procedure just presented the Routh array of coefficients is constructed. The first two rows can be obtained directly from the given polynomial. The remaining terms are obtained by the given schedule. (If any coefficients are missing, they are replaced by zeros in the array.)

$$\begin{array}{lllll}
 & \lambda^4 & 1 & 3 & 5 \\
 & \lambda^3 & 2 & 4 & 0 \\
\text{one sign change} & \lambda^2 & 1 & 5 & 0 \\
 & \lambda^1 & -6 & 0 & \\
\text{one sign change} & \lambda^0 & 5 & &
\end{array}$$

$$\begin{array}{llllll}
 & \lambda^4 & 1 & 3 & 5 & \\
 & \lambda^3 & 1 & 2 & 0 & \leftarrow \text{second row is divided by two} \\
\text{one sign change} & \lambda^2 & 1 & 5 & 0 & \\
 & \lambda^1 & -3 & 0 & & \\
\text{one sign change} & \lambda^0 & 5 & & &
\end{array}$$

In this example the number of changes of sign of the coefficients in the first column is two. This means that there are two roots with positive real parts. In this example it is demonstrated that the result is unchanged when the coefficients of any row are multiplied or divided by a positive number in order to simplify the evaluation of coefficients.

3. If the first column term in any row is zero but the remaining terms are not zero, the zero term is replaced by a very small positive number ϵ and the rest of the array is evaluated.

Example 8-3 Consider the following equation:

$$\lambda^3 + 2\lambda^2 + \lambda + 2 = 0 \tag{8-4}$$

The Routh array of coefficients is

λ^3	1	1
λ^2	2	2
λ^1	$0 \approx \epsilon$	0
λ^0	2	

If the sign of the coefficient above the zero (ϵ) is the same as that below it, it indicates that there are a pair of imaginary roots. Actually Eq. (8–4) has roots $\lambda = j$ and $\lambda = -j$. If, however, the sign of the coefficient above the zero (ϵ) is opposite that below it, it indicates that there is one sign change. For example, for the following equation:

$$\lambda^3 - 2\lambda^2 - \lambda + 2 = (\lambda + 1)(\lambda - 1)(\lambda - 2) = 0$$

the array of coefficients is

one sign change	λ^3	1	-1
	λ^2	-2	2
one sign change	λ^1	$0 \approx \epsilon$	0
	λ^0	2	

Thus, there are two sign changes of the coefficients in the first column, which agrees with the correct result indicated by the factored form of the polynomial equation.

4. If all the coefficients in any derived row are zero, this indicates a pair or pairs of real roots with equal magnitudes but opposite signs and/or a pair or pairs of conjugate imaginary roots. In such cases, evaluation of the rest of the array can be continued by forming an auxiliary polynomial with coefficients of the last row and taking the coefficients of the derivative of this polynomial for the next row. Such roots with equal magnitudes but opposite signs can be found by solving the auxiliary polynomial which is always even. For a $2n$th-degree auxiliary polynomial, there are n pairs of equal and opposite roots.

Example 8-4 Consider the following equation:

$$\lambda^5 + 2\lambda^4 + 24\lambda^3 + 48\lambda^2 - 25\lambda - 50 = 0$$

The Routh array of coefficients is

λ^5	1	24	-25
λ^4	2	48	-50 → auxiliary polynomial $P(\lambda)$
λ^3	0	0	

The terms in the λ^3 row are all zero. The auxiliary polynomial is then formed from the coefficients of the λ^4 row. The auxiliary polynomial $P(\lambda)$ is given by

$$P(\lambda) = 2\lambda^4 + 48\lambda^2 - 50$$

which indicates that there are two pairs of roots of equal magnitudes and opposite signs. These two pairs of roots are obtained by solving the auxiliary polynomial equation $P(\lambda) = 0$. The derivative of $P(\lambda)$ with respect to λ is obtained as

$$\frac{dP(\lambda)}{d\lambda} = 8\lambda^3 + 96\lambda$$

The terms in the λ^3 row are replaced by the coefficients of the last equation, namely, 8 and 96. The array of coefficients then becomes as follows:

	λ^5	1	24	-25
	λ^4	2	48	-50
	λ^3	8	96	0 ← coefficients of $\frac{dP(\lambda)}{d\lambda}$
	λ^2	24	-50	0
one sign change	λ^1	112.7	0	
	λ^0	-50		

It is seen that there is one change in sign in the first column of the new array. Thus, the original equation has one root with a positive real part. By solving for roots of the auxiliary polynomial equation,

$$2\lambda^4 + 48\lambda^2 - 50 = 0$$

we obtain

$$\lambda = \pm 1 \quad \text{or} \quad \lambda = \pm 5j$$

These two pairs of roots are part of the roots of the original equation. As a matter of fact the original equation can be written in a factored form as follows:

$$(\lambda + 1)(\lambda - 1)(\lambda + 5j)(\lambda - 5j)(\lambda + 2) = 0$$

Clearly, the original equation has one root with a positive real part.

5. Finally it is noted that if the following substitution is made in Eq. (8–3)

$$\lambda = \mu - \sigma$$

where σ is a constant and the polynomial is written in terms of μ, then the application of Routh's stability criterion reveals the number of

roots which have real parts greater than $-\sigma$. Namely the number of changes of sign in the first column of the Routh array developed for the polynomial in μ is equal to the number of roots which have real parts greater than $-\sigma$.

Hurwitz stability criterion Hurwitz stability criterion gives conditions for the coefficients of the polynomial given by Eq. (8–3) to be satisfied in order that all the roots have negative real parts. As stated in preceding discussions, in order that all the roots have negative real parts, all the coefficients a's must be positive. This is a necessary condition but not a sufficient condition. If this condition is not satisfied it indicates that some of the roots have positive real parts or are imaginary or zero. The sufficient condition for all the roots to have negative real parts is given in the following: If all the coefficients of the polynomial are positive, arrange these coefficients in the following determinant:

$$\Delta_n = \begin{vmatrix} a_1 & a_3 & a_5 & \cdots & 0 & 0 & 0 \\ a_0 & a_2 & a_4 & \cdots & \cdot & \cdot & \cdot \\ 0 & a_1 & a_3 & \cdots & a_n & 0 & 0 \\ 0 & a_0 & a_2 & \cdots & a_{n-1} & 0 & 0 \\ \cdot & \cdot & \cdot & & a_{n-2} & a_n & 0 \\ \cdot & \cdot & \cdot & & a_{n-3} & a_{n-1} & 0 \\ 0 & 0 & 0 & \cdots & a_{n-4} & a_{n-2} & a_n \end{vmatrix}$$

where we substituted zero for a_s if $s > n$, or alternatively,

$$\Delta_n = \begin{vmatrix} a_1 & a_0 & 0 & 0 & \cdot & \cdot & 0 \\ a_3 & a_2 & a_1 & a_0 & \cdot & \cdot & 0 \\ a_5 & a_4 & a_3 & a_2 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & \cdot & \cdot & & & \cdot \\ 0 & \cdot & a_n & a_{n-1} & a_{n-2} & a_{n-3} & a_{n-4} \\ 0 & \cdot & 0 & 0 & a_n & a_{n-1} & a_{n-2} \\ 0 & \cdot & 0 & 0 & 0 & 0 & a_n \end{vmatrix}$$

For all the roots to have negative real parts, it is necessary and sufficient that successive principal minors of Δ_n be positive. The successive principal minors are the following determinants:

$$\Delta_i = \begin{vmatrix} a_1 & a_3 & \cdots & a_{2i-1} \\ a_0 & a_2 & \cdots & a_{2i-2} \\ 0 & a_1 & \cdots & a_{2i-3} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & a_i \end{vmatrix} = \begin{vmatrix} a_1 & a_0 & \cdots & 0 \\ a_3 & a_2 & \cdots & 0 \\ a_5 & a_4 & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{2i-1} & a_{2i-2} & \cdots & a_i \end{vmatrix}$$

$$(i = 1, 2, \ldots, n)$$

where $a_s = 0$ if $s > n$. (It is noted that some of the conditions for the lower-order determinants are included in the conditions for the higher-order determinants.) If all these determinants are positive, and $a_0 > 0$ as already assumed, the equilibrium state of the system whose characteristic equation is given by Eq. (8–3) is asymptotically stable. Note that exact values of determinants are not needed, instead only signs of these determinants are needed for the stability criterion.

Example 8-5 For the following characteristic equation:

$$a_0\lambda^4 + a_1\lambda^3 + a_2\lambda^2 + a_3\lambda + a_4 = 0$$

the conditions for asymptotic stability are that all the a's be positive and that

$$\Delta_2 = \begin{vmatrix} a_1 & a_3 \\ a_0 & a_2 \end{vmatrix} = a_1a_2 - a_0a_3 > 0$$

$$\Delta_3 = \begin{vmatrix} a_1 & a_3 & 0 \\ a_0 & a_2 & a_4 \\ 0 & a_1 & a_3 \end{vmatrix}$$

$$= a_1(a_2a_3 - a_1a_4) - a_0a_3^2$$

$$= a_3(a_1a_2 - a_0a_3) - a_1^2a_4 > 0$$

It is clear that if all the a's are positive and if the condition $\Delta_3 > 0$ is satisfied, the condition $\Delta_2 > 0$ is also satisfied. Therefore, for all the roots of the given characteristic equation to have negative real parts, it is necessary and sufficient that all the coefficients a's are positive and $\Delta_3 > 0$.

Equivalence of Routh's stability criterion and Hurwitz stability criterion The equivalence of Routh's stability criterion and Hurwitz stability criterion can be easily established. If we write the Hurwitz determinants in a triangular form such that

$$\Delta_i = \begin{vmatrix} a_{11} & & & & * \\ & a_{22} & & & \\ & & a_{33} & & \\ & & & \ddots & \\ 0 & & & & a_{ii} \end{vmatrix} \qquad (i = 1, 2, \ldots, n)$$

where the elements in the lower left of the main diagonal are zero and those in the upper right of the main diagonal are zero or nonzero, then, clearly

$$\Delta_i = a_{11}a_{22} \cdots a_{ii} \qquad (i = 1, 2, \ldots, n)$$

Using this fact the Hurwitz conditions

$$\Delta_1 > 0, \quad \Delta_2 > 0, \quad \Delta_3 > 0, \quad \ldots, \quad \Delta_n > 0$$

reduce to the conditions

$$a_{11} > 0, \quad a_{22} > 0, \quad a_{33} > 0, \quad \ldots, \quad a_{nn} > 0$$

It can be shown without difficulty that in terms of a_1, b_1, c_1, and so on, in the Routh array, $a_{11}, a_{22}, a_{33}, \ldots$ can be obtained as

$$a_{11} = a_1, \quad a_{22} = b_1, \quad a_{33} = c_1, \quad \ldots$$

Hence the Hurwitz conditions for asymptotic stability become

$$a_1 > 0, \quad b_1 > 0, \quad c_1 > 0, \quad \ldots$$

which are the Routh conditions for asymptotic stability. (For details, see Prob. A-8-3) In this way the equivalence of Routh's stability criterion and Hurwitz stability criterion can be established.

8-3 FIRST METHOD OF LIAPUNOV (8-4)

The first method of Liapunov consists of all procedures in which the explicit form of the solution is used for stability analysis. In this method each equilibrium state, if there is more than one equilibrium state, is investigated separately.

Consider the nonlinear system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) \tag{8-5}$$

where $\mathbf{x}$ is a state vector (n vector), and $\mathbf{f}(\mathbf{x})$ is an n vector and is continuously differentiable in $x_1, x_2, \ldots, x_n$. Let us expand the nonlinear vector function $\mathbf{f}(\mathbf{x})$ in Taylor series about the equilibrium state $\mathbf{x}_e$ in question. Introducing a new vector, $\mathbf{y} = \mathbf{x} - \mathbf{x}_e$, the equilibrium state can be shifted to the origin. By expanding $\mathbf{f}(\mathbf{x})$ in Taylor series about $\mathbf{x} = \mathbf{x}_e$, Eq. (8-5) becomes

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y} + \mathbf{G}(\mathbf{y})\mathbf{y}$$

where $\mathbf{A}$ is the $n \times n$ Jacobian matrix given by

$$\mathbf{A} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \cdots & \dfrac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{bmatrix}$$

where $f_1, f_2, \ldots, f_n$ are the n components of $\mathbf{f}(\mathbf{x})$. All the partial derivatives appearing in the Jacobian matrix $\mathbf{A}$ are evaluated at the equilibrium state $\mathbf{x} = \mathbf{x}_e$, or $\mathbf{y} = \mathbf{0}$. The $n \times n$ matrix $\mathbf{G}(\mathbf{y})$ contains terms arising from the higher-order derivatives in the Taylor series expansion. The elements of $\mathbf{G}(\mathbf{y})$ vanish at the equilibrium state. Equation (8-5) can thus be linearized near the origin as follows:

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y} \tag{8-6}$$

Equation (8-6) is the first approximation to the nonlinear equation, Eq. (8-5). We say that Eq. (8-6) is obtained from Eq. (8-5) by *linearization*. Liapunov showed that if all the eigenvalues of the constant matrix $\mathbf{A}$ in Eq. (8-6) have nonzero real parts, then the stability of the equilibrium state $\mathbf{x}_e$ of the original nonlinear equation, Eq. (8-5), is the same as that of the equilibrium state $\mathbf{y} = \mathbf{0}$ of the linearized equation, Eq. (8-6). Hence if the eigenvalues of $\mathbf{A}$ have negative real parts, then the equilibrium state $\mathbf{x}_e$ is asymptotically stable and all solutions of the system given by Eq. (8-5), with initial state $\mathbf{x}(0)$ sufficiently close to $\mathbf{x}_e$, approach $\mathbf{x}_e$ as $t \to \infty$ and if at least one of the eigenvalues of $\mathbf{A}$ has a positive real part, then the equilibrium state $\mathbf{x}_e$ is unstable. If, however, at least one of the eigenvalues of $\mathbf{A}$ has a zero real part, the local stability behavior of the equilibrium state $\mathbf{x} = \mathbf{x}_e$ of the system, Eq. (8-5), cannot be determined by Eq. (8-6). We shall call this case the *critical case*. In the critical case the local stability behavior of the equilibrium state of the system depends on higher-order terms of the Taylor series expansion of $\mathbf{f}(\mathbf{x})$. For example, if one eigenvalue of the Jacobian matrix $\mathbf{A}$ is zero at the equilibrium state, then this state is asymptotically stable if in its neighborhood $\mathbf{A}$ has only eigenvalues with negative real parts. The equilibrium state is unstable, however, if in its neighborhood there is at least one eigenvalue with a positive real part.

The conclusions by the stability analysis in the first approximation apply only in the neighborhood of the equilibrium state. Namely, the first method of Liapunov concerns only stability in the small; that is, whether a motion originating from a neighborhood of an equilibrium state $\mathbf{x}_e$ will approach $\mathbf{x}_e$ as t increases indefinitely. The method gives no indication of how small

the neighborhood must be in order that the motion approach $\mathbf{x}_e$ as $t \to \infty$. Stability analysis in the large is more difficult than stability analysis in the small. The second method of Liapunov to be presented in section 8–4 answers some of the questions of stability analysis in the large as well as local stability analysis of linear and nonlinear systems.

8-4 SECOND METHOD OF LIAPUNOV[1]

The second method of Liapunov attempts to give information on the stability of equilibrium states of linear and nonlinear systems without any knowledge of their solutions. The stability information obtained by this method is precise and involves no approximation.

The second method of Liapunov is based on a generalization of the idea that if the system has an asymptotically stable equilibrium state, then the stored energy of the system displaced within the domain of attraction decays with increasing time until it finally assumes its minimum value at the equilibrium state. The second method of Liapunov consists of determination of a fictitious "energy" function called a *Liapunov function*. The idea of the Liapunov function is more general than that of energy and is more widely applicable. Liapunov functions are functions of $x_1, x_2, \ldots, x_n$, and t. We denote Liapunov functions as $V(x_1, x_2, \ldots, x_n, t)$ or simply as $V(\mathbf{x}, t)$. If Liapunov functions do not include t explicitly, then we denote them as $V(x_1, x_2, \ldots, x_n)$ or $V(\mathbf{x})$. Liapunov functions can sometimes be expressed as functions of not only $x_1, x_2, \ldots, x_n, t$, but also $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$, where the bar denotes the complex conjugate. In this book we consider only physical systems. In such systems complex elements of the state vector $\mathbf{x}$ must occur as complex conjugate pairs. Hence for such systems $V(x_1, x_2, \ldots, x_n; \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n, t)$ can be expressed simply as $V(x_1, x_2, \ldots, x_n, t)$ or $V(\mathbf{x}, t)$, because if x_i and x_j are a complex conjugate pair then $\bar{x}_i = x_j$, and if x_k is real then $\bar{x}_k = x_k$. In the second method of Liapunov the sign behaviors of $V(\mathbf{x}, t)$ and its time derivative $\dot{V}(\mathbf{x}, t)$ give information on stability, asymptotic stability or instability of the equilibrium state under consideration without directly solving for the solution. (This applies to both linear and nonlinear systems.)

In this section we first present Liapunov's main stability theorem and its proof and then present several modified forms of the stability theorem. We assume that the equilibrium state under consideration is at the origin of the state space. The essence of the second method of Liapunov is given in the following main theorem:

[1]This section follows closely (*8–2, 8–3*). Numbers in parentheses refer to titles in the references listed at the end of the chapter.

Theorem 8-1 (*Liapunov's main stability theorem*) Let the system be defined by Eq. (8–1) which is rewritten thus:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t)$$

Suppose that

$$\mathbf{f}(\mathbf{0}, t) = \mathbf{0} \qquad \text{for all } t$$

Suppose also that there exists a scalar function $V(\mathbf{x}, t)$ which has continuous first partial derivatives. If $V(\mathbf{x}, t)$ satisfies the following conditions:

1. $V(\mathbf{x}, t)$ is positive definite, namely, $V(\mathbf{0}, t) = 0$ and
2. $V(\mathbf{x}, t) \geqslant \alpha(\|\mathbf{x}\|) > 0$ for all $\mathbf{x} \neq \mathbf{0}$ and all t where α is a continuous, nondecreasing scalar function such that $\alpha(0) = 0$.
3. The total derivative $\dot{V}$ is negative for all $\mathbf{x} \neq \mathbf{0}$ and all t, or $\dot{V}(\mathbf{x}, t) \leqslant -\gamma(\|\mathbf{x}\|) < 0$ for all $\mathbf{x} \neq \mathbf{0}$ and all t where γ is a continuous nondecreasing scalar function such that $\gamma(0) = 0$.
4. There exists a continuous, nondecreasing scalar function such that $\beta(0) = 0$ and, for all t, $V(\mathbf{x}, t) \leqslant \beta(\|\mathbf{x}\|)$
5. $\alpha(\|\mathbf{x}\|)$ approaches infinity as $\|\mathbf{x}\|$ increases indefinitely, or

$$\alpha(\|\mathbf{x}\|) \to \infty \qquad \text{as } \|\mathbf{x}\| \to \infty$$

then the origin of the system, $\mathbf{x} = \mathbf{0}$, is uniformly asymptotically stable in the large.

Proof To prove uniform asymptotic stability in the large we need to prove the following:

1. The origin is uniformly stable.
2. Every solution is uniformly bounded.
3. Every solution converges to the origin as $t \to \infty$ uniformly in t_0 and $\|\mathbf{x}_0\| \leqslant \delta$ where δ is fixed but arbitrarily large; that is, given two real numbers $\delta > 0$ and $\mu > 0$ there is a real number $T(\mu, \delta)$ such that

$$\|\mathbf{x}_0\| \leqslant \delta$$

implies

$$\|\phi(t; \mathbf{x}_0, t_0)\| \leqslant \mu \qquad \text{for all } t \geqslant t_0 + T(\mu, \delta)$$

where $\phi(t; \mathbf{x}_0, t_0)$ is the solution to the given differential equation.

Since β is continuous and $\beta(0) = 0$, we can take $\delta(\epsilon) > 0$ such that $\beta(\delta) < \alpha(\epsilon)$ for any $\epsilon > 0$. Figure 8–4 shows curves $\alpha(\|\mathbf{x}\|)$, $\beta(\|\mathbf{x}\|)$, and $V(\mathbf{x}, t)$. Noting that

$$V(\phi(t; \mathbf{x}_0, t_0), t) - V(\mathbf{x}_0, t_0)$$
$$= \int_{t_0}^{t} \dot{V}(\phi(\tau; \mathbf{x}_0, t_0), \tau)\, d\tau < 0, \qquad t > t_0$$

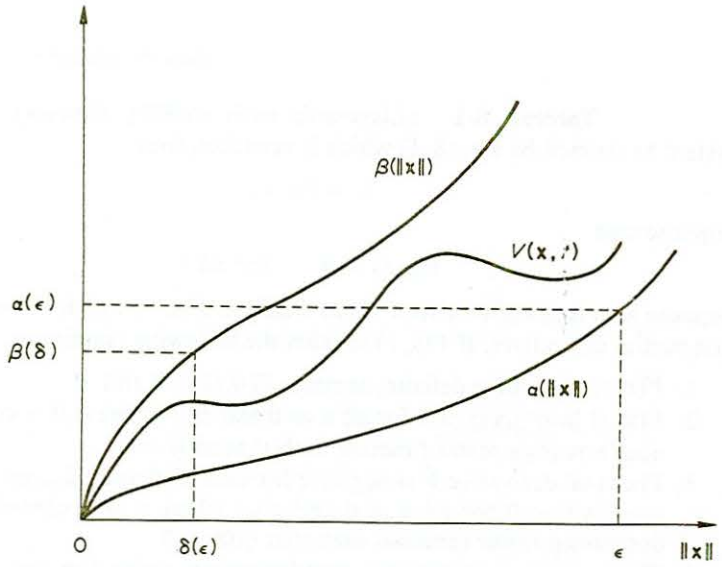


Figure 8-4 Curves $\alpha(\|\mathbf{x}\|)$, $\beta(\|\mathbf{x}\|)$, and $V(\mathbf{x}, t)$.

if $\|\mathbf{x}_0\| \leqslant \delta$, t_0 being aribtrary, we have

$$\alpha(\epsilon) > \beta(\delta) \geqslant V(\mathbf{x}_0, t_0) \geqslant V(\boldsymbol{\phi}(t; \mathbf{x}_0, t_0), t) \geqslant \alpha(\|\boldsymbol{\phi}(t; \mathbf{x}_0, t_0)\|)$$

for all $t \geqslant t_0$. Since α is nondecreasing and positive, this implies that

$$\|\boldsymbol{\phi}(t; \mathbf{x}_0, t_0)\| < \epsilon \qquad \text{for} \qquad t \geqslant t_0, \ \|\mathbf{x}_0\| \leqslant \delta$$

Hence we have shown that for each real number $\epsilon > 0$, there is a real number $\delta > 0$ such that $\|\mathbf{x}_0\| \leqslant \delta$ implies $\|\boldsymbol{\phi}(t; \mathbf{x}_0, t_0)\| \leqslant \epsilon$ for all $t \geqslant t_0$. Thus we have proved uniform stability.

Next, we shall prove that $\|\boldsymbol{\phi}(t; \mathbf{x}_0, t_0)\| \to 0$ with $t \to \infty$ uniformly in t_0 and $\|\mathbf{x}_0\| \leqslant \delta$. Let us take any $0 < \mu < \|\mathbf{x}_0\|$ and find a $\nu(\mu) > 0$ such that $\beta(\nu) < \alpha(\mu)$. Let us denote by $\epsilon'(\mu, \delta) > 0$ the minimum of the continuous nondecreasing function $\gamma(\|\mathbf{x}\|)$ on the compact set $\nu(\mu) \leqslant \|\mathbf{x}\| \leqslant \epsilon(\delta)$. Let us define

$$T(\mu, \delta) = \frac{\beta(\delta)}{\epsilon'(\mu, \delta)} > 0$$

Suppose that $\|\boldsymbol{\phi}(t; \mathbf{x}_0, t_0)\| > \nu$ over the time interval $t_0 \leqslant t \leqslant t_1 = t_0 + T$. Then, we have

$$0 < \alpha(\nu) \leqslant V(\boldsymbol{\phi}(t_1; \mathbf{x}_0, t_0), t_1) \leqslant V(\mathbf{x}_0, t_0) - (t_1 - t_0)\epsilon' \leqslant \beta(\delta) - T\epsilon' = 0$$

which is a contradiction. Hence for some t in the interval $t_0 \leqslant t \leqslant t_1$, say, t_2 we have

$$\|\mathbf{x}_2\| = \|\boldsymbol{\phi}(t_2; \mathbf{x}_0, t_0)\| = \nu$$

Therefore,

$$\alpha(\|\boldsymbol{\phi}(t; \mathbf{x}_2, t_2)\|) \leqslant V(\boldsymbol{\phi}(t; \mathbf{x}_2, t_2), t) \leqslant V(\mathbf{x}_2, t_2) \leqslant \beta(\nu) < \alpha(\mu)$$

for all $t \geqslant t_2$. Hence,

$$\|\boldsymbol{\phi}(t; \mathbf{x}_0, t_0)\| < \mu$$

for all $t \geqslant t_0 + T(\mu, \delta) \geqslant t_2$ which proves uniform asymptotic stability. Since $\alpha(\|\mathbf{x}\|) \to \infty$ as $\|\mathbf{x}\| \to \infty$ there exists for arbitrary large δ a constant $\epsilon(\delta)$ such that $\beta(\delta) < \alpha(\epsilon)$. Moreover, since $\epsilon(\delta)$ does not depend on t_0, the solution $\boldsymbol{\phi}(t; \mathbf{x}_0, t_0)$ is uniformly bounded. We thus have proved uniform asymptotic stability in the large.

Liapunov function The Liapunov function $V(\mathbf{x}, t)$, a scalar function, is a positive definite function, and it is continuous together with its first partial derivatives (with respect to its arguments) in the region Ω about the origin and has a time derivative which, when taken along the trajectory, is negative definite (or semidefinite). Notice that $\dot{V}(\mathbf{x}, t)$ is actually the total derivative of $V(\mathbf{x}, t)$ with respect to t along a solution of the system. Hence $\dot{V}(\mathbf{x}, t) < 0$ implies that $V(\mathbf{x}, t)$ is a decreasing function of t. A Liapunov function is not unique in a given system. (For this reason, the second method of Liapunov is a more powerful tool than conventional energy considerations. Note that a system whose energy E decreases on the average, but not necessarily at each instant, is stable but E is not a Liapunov function.)

The simplest positive definite function is of a quadratic form

$$V(\mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} q_{ij} x_i x_j \qquad (i, j = 1, 2, \ldots, n)$$

In general, Liapunov functions may not be of a simple quadratic form. For any Liapunov function, the lowest-degree terms in V must be even. This can be seen as follows: If we define

$$\frac{x_1}{x_n} = z_1, \qquad \frac{x_2}{x_n} = z_2, \qquad \ldots, \qquad \frac{x_{n-1}}{x_n} = z_{n-1}$$

then in the neighborhood of the origin only the lowest degree terms become dominant and we can write $V(\mathbf{x})$ as

$$V(\mathbf{x}) = x_n^p V(z_1, z_2, \ldots, z_{n-1}, 1)$$

If we keep the z_i fixed, $V(z_1, z_2, \ldots, z_{n-1}, 1)$ is a fixed quantity. For p odd, x_n^p can assume both positive and negative values near the origin, which means that $V(\mathbf{x})$ is not positive definite. Hence p must be even.

The second method of Liapunov is simple to apply in principle, but in

practice it is not necessarily easy for many cases, because no general method for constructing Liapunov functions is available at present, although there is a great flexibility in choosing a Liapunov function. There is, however, no need to solve the differential equation. It should be noted that failure to find a suitable V function does not imply instability but merely means that the stability condition cannot be guaranteed.

Example 8-6 (8-2) Consider the system described by

$$\begin{aligned} \dot{x}_1 &= x_2 - ax_1(x_1^2 + x_2^2) \\ \dot{x}_2 &= -x_1 - ax_2(x_1^2 + x_2^2) \end{aligned} \tag{8-7}$$

where a is a positive constant. Clearly the equilibrium state is the origin, or $\mathbf{x} = \mathbf{0}$. If we define a scalar function $V(\mathbf{x})$ of the state variables by

$$V(\mathbf{x}) = x_1^2 + x_2^2$$

which is positive except at the origin where it is zero, then the time derivative $\dot{V}(\mathbf{x})$ along a solution of Eq. (8-7) is

$$\dot{V}(\mathbf{x}) = \frac{dV(\mathbf{x})}{dt} = -2a(x_1^2 + x_2^2)^2$$

which is negative unless $x_1 = x_2 = 0$. This shows that $V(\mathbf{x})$ is constantly decreasing along any solution of Eq. (8-7), and hence $V(\mathbf{x})$ is a Liapunov function. Let $V(\mathbf{x})$ take constant values, $0, C_1, C_2, C_3, \ldots, (0 < C_1 < C_2 < C_3 \cdots)$. Then $V = 0$ corresponds to the origin of the state plane. $V = C_1$, $V = C_2, \ldots$ describe non-intersecting circles enclosing the origin of the state plane as shown in Fig. 8-5.

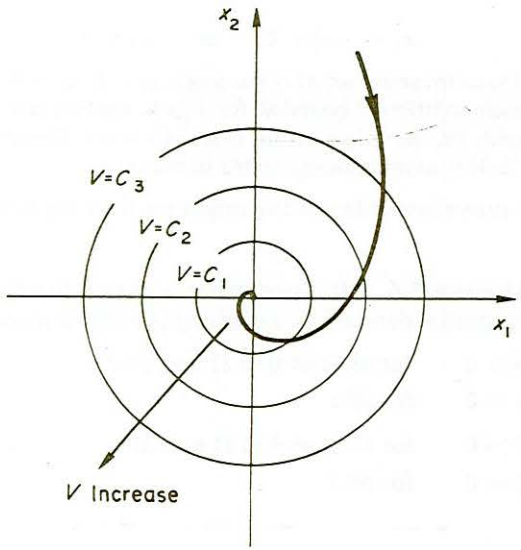


Figure 8-5 Constant V contours and representative trajectory.

The circle $V = C_k$ lies completely inside of the circle $V = C_{k+1}$. A representative trajectory crosses the boundary of V contours from outside toward inside. From this, the geometric interpretation of a Liapunov function may be stated as follows: $V(\mathbf{x})$ is a measure of the *distance* of the state $\mathbf{x}$ from the origin of the state space. If the distance between the origin and the instantaneous state $\mathbf{x}(t)$ is continually decreasing as t increases (that is, $\dot{V}(\mathbf{x}(t)) < 0$), then $\mathbf{x}(t) \to \mathbf{0}$. Since a Liapunov function $V(\mathbf{x})$ is regarded as defining *distance* from the origin in the state space, its derivative can be used to give a quantitative estimate of the speed with which the origin is approached. Note that in this example $V(\mathbf{x})$ becomes infinite with infinite deviation from the equilibrium state. Hence the equilibrium state, the origin, of the system is asymptotically stable for all possible initial conditions, or asymptotically stable in the large.

Several stability theorems which are based on Theorem 8–1 are now presented. The system differential equation considered in Theorems 8–2 to 8–6 is defined by

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t) \tag{8-8}$$

where $\mathbf{x}$ is a state vector (n vector) and $\mathbf{f}(\mathbf{x}, t)$ is an n vector having continuous first partial derivatives in the region Ω (Ω can be the entire state space) which includes the origin. We assume that $\mathbf{f}(\mathbf{0}, t) = \mathbf{0}$; namely, there is an equilibrium at the origin. We also assume that notation $V(\mathbf{x}, t) > 0$ implies that Conditions 2 and 4 of Theorem 8–1 are satisfied.

Theorem 8-2 If there exists a scalar function $V(\mathbf{x}, t)$, with continuous first partial derivatives, satisfying the following conditions:

(a) $V(\mathbf{x}, t) > 0$ for all $\mathbf{x} \neq \mathbf{0}$ in Ω and all t
$V(\mathbf{0}, t) = 0$ for all t

(b) $\dot{V}(\mathbf{x}, t) < 0$ for all $\mathbf{x} \neq \mathbf{0}$ in Ω and all t
$\dot{V}(\mathbf{0}, t) = 0$ for all t

then the origin of the system of Eq. (8-8) is uniformly asymptotically stable.

If the foregoing conditions are satisfied in the whole state space and either $V(\mathbf{x}, t) \to \infty$ as $\|\mathbf{x}\| \to \infty$, or $\dot{V}(\mathbf{x}, t) \leqslant -m < 0$ for all $\|\mathbf{x}\| \geqslant M$ and some m, $M > 0$, then the origin is uniformly asymptotically stable in the large.

If asymptotic stability is assured only for all possible motions starting within Ω, then practical considerations may determine whether the size of Ω is sufficient.

To prove stability (but not asymptotic stability) of the origin the following theorem can be applied:

Theorem 8-3 If there exists a scalar function $V(\mathbf{x}, t)$, with continuous first partial derivatives, satisfying the following conditions:

(a) $V(\mathbf{x}, t) > 0$ for all $\mathbf{x} \neq \mathbf{0}$ in Ω and all t
$V(\mathbf{0}, t) = 0$ for all t

(b) $\dot{V}(\mathbf{x}, t) \leqslant 0$ for all $\mathbf{x} \neq \mathbf{0}$ in Ω and all t
$\dot{V}(\mathbf{0}, t) = 0$ for all t

then the origin of the system of Eq. (8–8) is uniformly stable.

This theorem states that the existence of such a $V(\mathbf{x}, t)$ function implies the uniform stability of the origin. It should be noted that $\dot{V}(\mathbf{x}, t) \leqslant 0$ along the trajectories means that the origin is uniformly stable but not necessarily uniformly asymptotically stable. Hence, in this case the system may exhibit a limit cycle operation.

Theorem 8-4 Suppose that there exists a scalar function $V(\mathbf{x}, t)$, with continuous first partial derivatives, satisfying the following conditions:

(a) $V(\mathbf{x}, t) > 0$ for all $\mathbf{x} \neq \mathbf{0}$ and all t
$V(\mathbf{0}, t) = 0$ for all t

(b) $\dot{V}(\mathbf{x}, t) \leqslant 0$ for all $\mathbf{x} \neq \mathbf{0}$ and all t
$\dot{V}(\mathbf{0}, t) = 0$ for all t

Let E be the set of all points where $\dot{V}(\mathbf{x}, t) = 0$, and let M be the maximum invariant set contained in E. (A set M is said to be *invariant* if each trajectory originating in M remains in M for all t.) Then every solution of Eq. (8–8) bounded for $t \geqslant 0$ approaches M as $t \to \infty$. If, in addition, $V(\mathbf{x}, t) \to \infty$ as $\|\mathbf{x}\| \to \infty$, then each solution is bounded for $t \geqslant 0$, and all solutions approach M as $t \to \infty$.

We say that $\mathbf{x}(t)$ approaches a set M as t approaches infinity if for each $\epsilon > 0$ there is a $T > 0$ with the property that, for each $t > T$ there is a state $\mathbf{p}$ in M with $\|\mathbf{x}(t) - \mathbf{p}\| < \epsilon$; that is, for all $t > T$ the states $\mathbf{x}(t)$ are within a distance ϵ of M. If the maximum invariant set contained in E is the origin, $\mathbf{x} = \mathbf{0}$, then we may state this theorem as follows:

Theorem 8-5 If there exists a scalar function $V(\mathbf{x}, t)$, with continuous first partial derivatives, satisfying the following conditions:

(a) $V(\mathbf{x}, t) > 0$ for all $\mathbf{x} \neq \mathbf{0}$ and all t
$V(\mathbf{0}, t) = 0$ for all t

(b) $\dot{V}(\mathbf{x}, t) \leqslant 0$ for all $\mathbf{x} \neq \mathbf{0}$ and all t
$\dot{V}(\mathbf{0}, t) = 0$ for all t

(c) $\dot{V}(\phi(t; \mathbf{x}_0, t_0), t)$ does not vanish identically in $t \geqslant t_0$ for any t_0 and any $\mathbf{x}_0 \neq \mathbf{0}$, where $\phi(t; \mathbf{x}_0, t_0)$ denotes the solution starting from $\mathbf{x}_0$ at t_0

then the origin of the system of Eq. (8–8) is uniformly asymptotically stable in the large.

This may be seen as follows: If $\dot{V}(\mathbf{x}, t)$ is not negative definite but only negative semidefinite, then the trajectory of the representative point can become tangent to some particular surface $V(\mathbf{x}, t) = C$. Since $\dot{V}(\phi(t; \mathbf{x}_0, t_0), t)$ does not vanish identically in $t \geqslant t_0$ for any t_0 and any $\mathbf{x}_0 \neq \mathbf{0}$, the representative point cannot remain at the tangent point [the point which corresponds to $\dot{V}(\mathbf{x}, t) = 0$] and therefore must move toward the origin.

Example 8-7 Consider the following system:

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_2 - x_1^3 \end{aligned} \tag{8-9}$$

Clearly the origin is the equilibrium state. Let us consider the following $V(x_1, x_2)$:

$$V(x_1, x_2) = x_1^4 + 2x_2^2$$

Then

$$\begin{aligned} \dot{V}(x_1, x_2) &= 4x_1^3 x_2 + 4x_2(-x_2 - x_1^3) \\ &= -4x_2^2 \leqslant 0 \end{aligned}$$

Let E be the set of all points where $\dot{V}(x_1, x_2) = 0$ and let M be the maximum invariant set contained in E,

$$\begin{aligned} E &= \{(x_1, x_2) \ ; \ \dot{V}(x_1, x_2) = 0\} \\ &= \{(x_1, 0) \ ; \ -\infty < x_1 < +\infty\} \end{aligned}$$

If

$$x_2 = 0$$

then

$$\dot{x}_2 = -x_1^3 = 0 \qquad \text{or} \qquad x_1 = 0$$

Hence the maximum invariant set M is the origin $x_1 = 0$, $x_2 = 0$. Since $V(\mathbf{x}) \to \infty$ as $\|\mathbf{x}\| \to \infty$, each solution is bounded for $t \geqslant 0$, and we can conclude that all solutions approach M, the origin in this case, as $t \to \infty$. Hence the origin of the system of Eq. (8–9) is asymptotically stable in the large.

Finally, to prove instability of the origin we have the following theorem:

Theorem 8-6 If there exists a scalar function $V(\mathbf{x}, t)$, with continuous first partial derivatives, satisfying the following conditions:

(a) $V(\mathbf{x}, t) > 0$ for all $\mathbf{x} \neq \mathbf{0}$ in Ω and all t
$V(\mathbf{0}, t) = 0$ for all t

(b) $\dot{V}(\mathbf{x}, t) > 0$ for all $\mathbf{x} \neq \mathbf{0}$ in Ω and all t
$\dot{V}(\mathbf{0}, t) = 0$ for all t

or

(a′) $V(\mathbf{x}, t) < 0$ for all $\mathbf{x} \neq \mathbf{0}$ in Ω and all t
$V(\mathbf{0}, t) = 0$ for all t

(b′) $\dot{V}(\mathbf{x}, t) < 0$ for all $\mathbf{x} \neq \mathbf{0}$ in Ω and all t
$\dot{V}(\mathbf{0}, t) = 0$ for all t

then the origin of the system of Eq. (8–8) is unstable.

This theorem states that if both $V(\mathbf{x}, t)$ and $\dot{V}(\mathbf{x}, t)$ are definite and have the same sign, then the origin is unstable and the trajectories near the origin diverge to a limiting value or infinity.

In concluding this section, note that in applying these stability theorems to a given system the stability conditions obtained from a particular V function are sufficient but may not be necessary. Furthermore, a Liapunov function for a particular system is not unique. Therefore, one should note the following:

1. Failure in finding a V function to show stability or asymptotic stability or instability of the equilibrium state under consideration can give no information on stability.
2. Although a particular V function may prove that the equilibrium state under consideration is stable or asymptotically stable in the region Ω which includes this equilibrium state, it does not necessarily mean that the motions are unstable outside the region Ω.
3. For a stable or asymptotically stable equilibrium state, a V function with the required properties always exists.

8-5 STABILITY ANALYSIS OF LINEAR SYSTEMS

In this section we first present applications of the second method of Liapunov to stability analysis of linear time-invariant systems. Then we give a proof of Routh-Hurwitz stability criteria. Finally, we present discussions of linear forced dynamic systems.

Liapunov stability analysis of linear time-invariant systems Consider the following system:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} \tag{8–10}$$

where $\mathbf{x}$ is a state vector (n vector) and $\mathbf{A}$ is an $n \times n$ constant matrix. We assume henceforth that $\mathbf{A}$ is nonsingular; that is, its determinant $|\mathbf{A}|$ is nonzero. The stability of the equilibrium state of the linear time-invariant system can be investigated easily by use of the second method of Liapunov.

The following is a basic theorem of stability analysis of linear time-invariant systems by means of the second method of Liapunov.

Theorem 8-7 The equilibrium state $\mathbf{x} = \mathbf{0}$ of the system given by Eq. (8–10) is asymptotically stable if and only if given any positive definite Hermitian matrix $\mathbf{Q}$ (or positive definite real symmetric matrix $\mathbf{Q}$), there exists a positive definite Hermitian matrix $\mathbf{P}$ (or positive definite real symmetric matrix $\mathbf{P}$) such that

$$\mathbf{A}^*\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{Q}$$

The scalar function $\mathbf{x}^*\mathbf{P}\mathbf{x}$ is a Liapunov function for the system of Eq. (8–10).

Proof We shall first prove that if there exists a positive definite Hermitian matrix $\mathbf{P}$ such that $\mathbf{A}^*\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{Q}$, then the equilibrium state of the system is asymptotically stable. Define

$$V(\mathbf{x}) = \mathbf{x}^*\mathbf{P}\mathbf{x} > 0 \qquad \text{for all } \mathbf{x} \neq \mathbf{0}$$

as a tentative Liapunov function for the system of Eq. (8–10). Then $\dot{V}$ along any trajectory is obtained as

$$\begin{aligned}\dot{V}(\mathbf{x}) &= \dot{\mathbf{x}}^*\mathbf{P}\mathbf{x} + \mathbf{x}^*\mathbf{P}\dot{\mathbf{x}} \\ &= \mathbf{x}^*\mathbf{A}^*\mathbf{P}\mathbf{x} + \mathbf{x}^*\mathbf{P}\mathbf{A}\mathbf{x} \\ &= \mathbf{x}^*(\mathbf{A}^*\mathbf{P} + \mathbf{P}\mathbf{A})\mathbf{x} \\ &= -\mathbf{x}^*\mathbf{Q}\mathbf{x}\end{aligned}$$

Thus

$$\dot{V}(\mathbf{x}) < 0 \qquad \text{for all } \mathbf{x} \neq \mathbf{0}$$

Hence the assumed $V(\mathbf{x})$ is a Liapunov function and the origin of the system is asymptotically stable.

We shall next prove that if the equilibrium state $\mathbf{x} = \mathbf{0}$ of the system is asymptotically stable, then there exists a positive definite Hermitian matrix $\mathbf{P}$ such that

$$\mathbf{A}^*\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{Q}$$

Referring to Prob. A-6-17, the matrix differential equation

$$\dot{\mathbf{X}} = \mathbf{A}^*\mathbf{X} + \mathbf{X}\mathbf{A}, \qquad \mathbf{X}(0) = \mathbf{Q} \tag{8–11}$$

has the solution

$$\mathbf{X} = e^{\mathbf{A}^* t}\mathbf{Q}e^{\mathbf{A}t}$$

Integrating both sides of Eq. (8–11) from $t = 0$ to $t = \infty$,

$$\mathbf{X}(\infty) - \mathbf{X}(0) = \mathbf{A}^*\left(\int_0^\infty \mathbf{X}\,dt\right) + \left(\int_0^\infty \mathbf{X}\,dt\right)\mathbf{A}$$

and noting that $\mathbf{A}$ is a stable matrix and, therefore, $\mathbf{X}(\infty) = \mathbf{0}$ we obtain

$$-\mathbf{Q} = \mathbf{A}^*\left(\int_0^\infty \mathbf{X}\,dt\right) + \left(\int_0^\infty \mathbf{X}\,dt\right)\mathbf{A}$$

Let us put

$$\mathbf{P} = \int_0^\infty \mathbf{X}\, dt = \int_0^\infty e^{\mathbf{A}^* t}\mathbf{Q}e^{\mathbf{A}t}\, dt$$

Note that the elements of $e^{\mathbf{A}t}$ are finite sums of terms like $e^{\lambda_i t}$, $te^{\lambda_i t}$, $\ldots, t^{m_i-1}e^{\lambda_i t}$ where the λ_i are the eigenvalues of $\mathbf{A}$ and m_i is the multiplicity of λ_i. Since the λ_i possess negative real parts,

$$\int_0^\infty e^{\mathbf{A}^* t}\mathbf{Q}e^{\mathbf{A}t}\, dt$$

exists. Note that

$$\mathbf{P}^* = \int_0^\infty e^{\mathbf{A}^* t}\mathbf{Q}e^{\mathbf{A}t}\, dt = \mathbf{P}$$

Thus, $\mathbf{P}$ is Hermitian (or symmetric if $\mathbf{P}$ is a real matrix). We have thus shown that for a stable $\mathbf{A}$ and for a positive definite Hermitian matrix $\mathbf{Q}$ there exists a Hermitian matrix $\mathbf{P}$ such that $\mathbf{A}^*\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{Q}$. We now need to prove that $\mathbf{P}$ is positive definite. Consider the following Hermitian form:

$$\mathbf{x}^*\mathbf{P}\mathbf{x} = \mathbf{x}^* \int_0^\infty e^{\mathbf{A}^* t}\mathbf{Q}e^{\mathbf{A}t}\, dt\, \mathbf{x}$$

$$= \int_0^\infty (e^{\mathbf{A}t}\mathbf{x})^*\mathbf{Q}(e^{\mathbf{A}t}\mathbf{x})\, dt > 0 \qquad \text{for } \mathbf{x} \neq \mathbf{0}$$

$$= 0 \qquad \text{for } \mathbf{x} = \mathbf{0}$$

Hence $\mathbf{P}$ is positive definite. This completes the proof.

In applying this theorem for stability analysis of linear time-invariant systems, we compute a Hermitian matrix $\mathbf{P}$ (or real symmetric matrix $\mathbf{P}$) by equating the matrices

$$\mathbf{A}^*\mathbf{P} + \mathbf{P}\mathbf{A} \qquad \text{and} \qquad -\mathbf{Q}$$

element by element. This results in $n(n+1)/2$ linear equations for determination of the elements $p_{ij} = \bar{p}_{ji}$ of $\mathbf{P}$. These simultaneous equations are soluble unless an eigenvalue of $\mathbf{A}$ or the sum of a pair of eigenvalues of $\mathbf{A}$ is zero. In other words if we define the eigenvalues of $\mathbf{A}$ as $\lambda_1, \lambda_2, \lambda_3, \ldots$, each repeated as often as its multiplicity as root of the characteristic equation, and if for every sum of the two roots, $\lambda_j + \lambda_k \neq 0$, then the matrix $\mathbf{P}$ is uniquely determined by the matrix $\mathbf{Q}$. Note that if the matrix $\mathbf{A}$ is a stable matrix then the sums $\lambda_j + \lambda_k$ are always negative. By applying Sylvester's criterion for $\mathbf{x}^*\mathbf{P}\mathbf{x}$ to be positive definite; that is,

$$p_{11} > 0, \quad \begin{vmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{vmatrix} > 0, \quad \begin{vmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{vmatrix} > 0, \quad \ldots$$

$$(p_{ij} = \bar{p}_{ji})$$

we can determine whether the origin of the system given by Eq. (8–10) is asymptotically stable. These conditions are clearly alternatives to the Routh or Hurwitz stability conditions.

Asymptotic stability can always be examined for a linear time-invariant system by solving a set of $n(n+1)/2$ simultaneous algebraic equations for the determination of a Liapunov function. (In the case of linear time-varying systems or nonlinear systems, however, no straightforward methods are presently available for finding Liapunov functions.)

One alternate form of Theorem 8–7 follows:

Theorem 8-8 A necessary and sufficient condition for $\mathbf{x} = \mathbf{0}$ to be an asymptotically stable solution of Eq. (8–10) is that there exist a positive definite Hermitian matrix $\mathbf{P}$ (or positive definite real symmetric matrix $\mathbf{P}$) satisfying the equation

$$\mathbf{A}^*\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{I}$$

where $\mathbf{I}$ is the identity matrix.

Example 8-8 Determine the stability of the equilibrium state of the following system:

$$\dot{x}_1 = -x_1 - 2x_2$$
$$\dot{x}_2 = x_1 - 4x_2$$

The equilibrium state is the origin or $\mathbf{x} = \mathbf{0}$. Let us first solve the following equation:

$$\mathbf{A}'\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{I}$$

or

$$\begin{bmatrix} -1 & 1 \\ -2 & -4 \end{bmatrix}\begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} + \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix}\begin{bmatrix} -1 & -2 \\ 1 & -4 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \tag{8-12}$$

where we have used the fact that $p_{21} = p_{12}$. If the matrix $\mathbf{P}$ turns out to be positive definite, then $\mathbf{x}'\mathbf{P}\mathbf{x}$ is a Liapunov function and the origin is asymptotically stable. Equation (8–12) yields the following three equations:

$$-2p_{11} + 2p_{12} = -1$$
$$-2p_{11} - 5p_{12} + p_{22} = 0$$
$$-4p_{12} - 8p_{22} = -1$$

Solving for p's, we obtain

$$p_{11} = \frac{23}{60}, \qquad p_{12} = -\frac{7}{60}, \qquad p_{22} = \frac{11}{60}$$

Hence

$$\mathbf{P} = \begin{bmatrix} \frac{23}{60} & -\frac{7}{60} \\ -\frac{7}{60} & \frac{11}{60} \end{bmatrix}$$

By Sylvester's criterion, this matrix is positive definite. Hence we conclude that the origin of the system is asymptotically stable. A Liapunov function is

$$V(\mathbf{x}) = \mathbf{x}'\mathbf{P}\mathbf{x} = [x_1 \quad x_2]\begin{bmatrix} \frac{23}{60} & -\frac{7}{60} \\ -\frac{7}{60} & \frac{11}{60} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{60}(23x_1^2 - 14x_1x_2 + 11x_2^2)$$

and $\dot{V}$ is given by

$$\dot{V} = -x_1^2 - x_2^2$$

Notice that in the asymptotically stable linear time-invariant system such as this, if $\mathbf{Q}$ is given to be a positive definite Hermitian matrix, then there exists a positive definite Hermitian matrix $\mathbf{P}$. For example, if in this system $\mathbf{Q}$ is given as

$$\mathbf{Q} = \begin{bmatrix} 1 & -1+j \\ -1-j & 4 \end{bmatrix}$$

then the matrix $\mathbf{P}$ that satisfies $\mathbf{A}^*\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{Q}$ is obtained as

$$\mathbf{P} = \begin{bmatrix} \frac{3}{10} & -\frac{1}{5} + j\frac{1}{5} \\ -\frac{1}{5} - j\frac{1}{5} & \frac{3}{5} \end{bmatrix}$$

$\mathbf{P}$ is clearly positive definite.

We shall next demonstrate that the following vector matrix differential equation: *(8–5)*

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} \tag{8–13}$$

where $\mathbf{x}$ is an n vector and $\mathbf{A}$ is an $n \times n$ constant matrix given by

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -b_n & 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -b_{n-1} & 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -b_3 & 0 & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -b_2 & -b_1 \end{bmatrix}$$

is equivalent to the following scalar differential equation:

$$\frac{d^n x}{dt^n} + a_1\frac{d^{n-1}x}{dt^{n-1}} + \cdots + a_n x = 0 \tag{8–14}$$

where $a_1, \ldots, a_n$ are real quantities, provided that a's and b's are related by

$$b_1 = \Delta_1, \qquad b_2 = \frac{\Delta_2}{\Delta_1}, \qquad b_3 = \frac{\Delta_3}{\Delta_1\Delta_2}, \qquad \ldots, \qquad b_i = \frac{\Delta_{i-3}\Delta_i}{\Delta_{i-2}\Delta_{i-1}} \tag{8–15}$$
$$(i = 4, 5, \ldots, n)$$

where $\Delta_1, \Delta_2, \ldots, \Delta_n$ are Hurwitz determinants

$$\Delta_i = \begin{bmatrix} a_1 & 1 & 0 & 0 & \cdots & 0 \\ a_3 & a_2 & a_1 & 1 & \cdots & 0 \\ a_5 & a_4 & a_3 & a_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ a_{2i-1} & a_{2i-2} & \cdot & \cdot & \cdots & a_i \end{bmatrix} \qquad (i = 1, 2, \ldots, n)$$

(a_s is replaced by zero if $s > n$)

and $\Delta_1, \Delta_2, \ldots, \Delta_{n-1}$ are nonzero quantities, and provided further that $x = x_1$. As an example, we shall demonstrate that the following equation:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} \tag{8–16}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ -b_3 & 0 & 1 \\ 0 & -b_2 & -b_1 \end{bmatrix}$$

is equivalent to

$$\frac{d^3x_1}{dt^3} + a_1\frac{d^2x_1}{dt^2} + a_2\frac{dx_1}{dx} + a_3x_1 = 0 \tag{8–17}$$

where a_1, a_2, and a_3, which are real quantities, are related to b_1, b_2, and b_3 by Eq. (8–15). Since

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ -b_3 & 0 & 1 \\ 0 & -b_2 & -b_1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_2 \\ -b_3x_1 + x_3 \\ -b_2x_2 - b_1x_3 \end{bmatrix}$$

we obtain

$$\dddot{x}_1 + b_1\ddot{x}_1 + (b_2 + b_3)\dot{x}_1 + b_1b_3x_1 = 0$$

Since b's are

$$b_1 = \Delta_1, \qquad b_2 = \frac{\Delta_2}{\Delta_1}, \qquad b_3 = \frac{\Delta_3}{\Delta_1\Delta_2}$$

where

$$\Delta_1 = a_1, \qquad \Delta_2 = \begin{vmatrix} a_1 & 1 \\ a_3 & a_2 \end{vmatrix} = a_1a_2 - a_3$$

$$\Delta_3 = \begin{vmatrix} a_1 & 1 & 0 \\ a_3 & a_2 & a_1 \\ 0 & 0 & a_3 \end{vmatrix} = a_1a_2a_3 - a_3^2$$

or

$$b_1 = a_1, \qquad b_2 = \frac{a_1 a_2 - a_3}{a_1}, \qquad b_3 = \frac{a_3}{a_1}$$

where we assumed $\Delta_1 \Delta_2 = a_1 (a_1 a_2 - a_3) \neq 0$, by substitution of these values of b's into the preceding third order scalar differential equation, we obtain

$$\dddot{x}_1 + a_1 \ddot{x}_1 + a_2 \dot{x}_1 + a_3 x_1 = 0$$

From this we see that if $a_1 \neq 0$, $a_1 a_2 \neq a_3$, then Eq. (8–16) can be modified into Eq. (8–17), and vice versa.

The following theorem essentially proves Hurwitz stability criterion. (Since Hurwitz stability criterion and Routh's stability criterion are equivalent, the following theorem also proves Routh's stability criterion.)

Theorem 8-9 (*8–2*) Consider the linear time-invariant system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} \tag{8–18}$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -b_n & 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -b_{n-1} & 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -b_3 & 0 & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -b_2 & -b_1 \end{bmatrix} \tag{8–19}$$

and $b_1, b_2, \ldots, b_n$ are real quantities. The origin of the system is asymptotically stable if and only if

$$b_1 > 0, \qquad b_2 > 0, \qquad \ldots, \qquad b_n > 0$$

(**A** is called *Schwarz matrix*.)

Proof Let us consider the scalar function $V = \mathbf{x}'\mathbf{P}\mathbf{x}$ as a possible Liapunov function, where

$$\mathbf{P} = \begin{bmatrix} b_1 b_2 \cdots b_n & 0 & \cdots & 0 & 0 & 0 \\ 0 & b_1 b_2 \cdots b_{n-1} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & b_1 b_2 b_3 & 0 & 0 \\ 0 & 0 & \cdots & 0 & b_1 b_2 & 0 \\ 0 & 0 & \cdots & 0 & 0 & b_1 \end{bmatrix}$$

By a simple calculation, we obtain

$$\dot{V} = \mathbf{x}'(\mathbf{A}'\mathbf{P} + \mathbf{P}\mathbf{A})\mathbf{x} = -2b_1^2 x_n^2$$

which is negative semidefinite. By examining Eq. (8–18) we find $\dot{V}$ cannot be identically equal to zero unless $x_i = 0$ $(i = 1, 2, \ldots, n)$. Hence the necessary and sufficient condition for the asymptotic stability of the origin is that $\mathbf{x}'\mathbf{P}\mathbf{x}$ be positive definite. This means that all the elements in the main diagonal of **P** must be positive; hence $b_i > 0$ $(i = 1, 2, \ldots, n)$. This completes the proof.

If we choose b's according to Eq. (8–15), then Eq. (8–18) is equivalent to Eq. (8–14). For asymptotic stability of the origin of the system of Eq. (8–18) or Eq. (8–14), Theorem 8–9 states that b's must be positive which, by Eq. (8–15), is equivalent to saying that Δ_i $(i = 1, 2, \ldots, n)$ must be positive. This proves Hurwitz stability criterion.

It is noted that we can obtain the correlation between the Routh array of the following equation:

$$(-1)^n |\mathbf{A} - \lambda \mathbf{I}| = 0$$

where **A** is given by Eq. (8–19) and the coefficients of powers of λ in successive principal minors of the determinant

$$|\mathbf{A} - \lambda \mathbf{I}|$$

starting from the top left-hand corner. By comparing the recurrence relationship involved between principal minors of orders i, $i - 1$, and $i - 2$ and the recurrence relationship defining the Routh array, we can prove the following general result which is stated as a theorem:

Theorem 8-10 (*8–5*) The first column of the Routh array of

$$(-1)^n |\mathbf{A} - \lambda \mathbf{I}| = 0$$

where **A** is given by Eq. (8–19) consists of the elements

$$1, \quad b_1, \quad b_2, \quad b_1 b_3, \quad b_2 b_4, \quad b_1 b_3 b_5, \quad b_2 b_4 b_6, \quad b_1 b_3 b_5 b_7, \quad \ldots$$

the final $(n + 1)$th element ending in b_n.[1]

We shall next present a theorem due to H. R. Schwarz. The proof is left to the reader. (Note that by use of Theorems 8–9, 8–10, and 8–11 we can immediately prove Routh's stability criterion.)

Theorem 8-11 (*8–2*) If a real constant matrix **C** is similar to the following matrix **A**

[1] See Prob. A-8-4.

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ -b_n & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -b_{n-1} & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & 0 & \cdots & -b_2 & -b_1 \end{bmatrix}$$

that is $\mathbf{C} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}$ ($\mathbf{T}$: nonsingular), then the number of eigenvalues of $\mathbf{C}$ which have negative real parts is equal to the number of positive terms in the sequence $b_1, b_1b_2, \ldots, b_1b_2 \cdots b_n$, provided that none of the b_i is zero.

Example 8-9 Consider the following matrix $\mathbf{C}$:

$$\mathbf{C} = \begin{bmatrix} -6 & -1 & -5 \\ 0 & 3 & 1 \\ 6 & 1 & 6 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 6 & 0 & 1 \\ 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$
$$= \mathbf{T}^{-1}\mathbf{A}\mathbf{T}$$

$\mathbf{C}$ is clearly similar to $\mathbf{A}$ where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 6 & 0 & 1 \\ 0 & 1 & 3 \end{bmatrix}$$

(Note that $\mathbf{C}$ and $\mathbf{A}$ have the same eigenvalues.) Since

$$b_1 = -3, \qquad b_2 = -1, \qquad b_3 = -6$$

the sequence

$$b_1 = -3$$
$$b_1b_2 = 3$$
$$b_1b_2b_3 = -18$$

has one positive term. Hence the number of eigenvalues of $\mathbf{C}$ (or $\mathbf{A}$) which have negative real parts is one. In fact, the characteristic equation of $\mathbf{C}$ (or $\mathbf{A}$) is given by

$$\lambda^3 - 3\lambda^2 - 7\lambda + 18 = 0$$

which has three roots

$$\lambda_1 = 2, \qquad \lambda_2 = 3.541, \qquad \lambda_3 = -2.541$$

Hence we can confirm that only one eigenvalue has a negative real part.

Linear forced dynamic systems Finally we shall present a theorem on the linear forced dynamic system.

Theorem 8-12 Consider the following system:

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u}$$

where $\mathbf{x} = n$ vector (state vector)

$\mathbf{u} = r$ vector (control vector or forcing vector)

$\mathbf{A}(t) = n \times n$ matrix

$\mathbf{B}(t) = n \times r$ matrix

If the fundamental matrix $\Phi(t, t_0)$ of

$$\dot{\Phi}(t, t_0) = \mathbf{A}(t)\Phi(t, t_0), \qquad \Phi(t_0, t_0) = \mathbf{I}$$

is bounded and also

$$\lim_{t\to\infty} \int_{t_0}^{t} \|\Phi(t, t_0)\Phi^{-1}(t_1, t_0)\mathbf{B}(t_1)\|\, dt_1$$

is bounded, then all solutions $\mathbf{x}$ for bounded $\mathbf{u}$ are bounded.

Proof The solution $\mathbf{x}(t)$ can be written as

$$\mathbf{x}(t) = \Phi(t, t_0)\mathbf{x}(t_0) + \Phi(t, t_0)\int_{t_0}^{t} \Phi^{-1}(t_1, t_0)\mathbf{B}(t_1)\mathbf{u}(t_1)\, dt_1$$

it follows that

$$\begin{aligned} \|\mathbf{x}(t)\| &\leqslant \|\Phi(t, t_0)\|\,\|\mathbf{x}(t_0)\| \\ &\quad + \int_{t_0}^{t} \|\Phi(t, t_0)\Phi^{-1}(t_1, t_0)\mathbf{B}(t_1)\|\,\|\mathbf{u}(t_1)\|\, dt_1 \\ &\leqslant \|\Phi(t, t_0)\|\,\|\mathbf{x}(t_0)\| \\ &\quad + \lim_{t\to\infty} \int_{t_0}^{t} \|\Phi(t, t_0)\Phi^{-1}(t_1, t_0)\mathbf{B}(t_1)\|\,\|\mathbf{u}(t_1)\|\, dt_1 \end{aligned}$$

Since $\mathbf{u}(t)$, $\|\Phi(t, t_0)\|$, and $\lim_{t\to\infty} \int_{t_0}^{t} \|\Phi(t, t_0)\Phi^{-1}(t_1, t_0)\mathbf{B}(t_1)\|\, dt_1$ are bounded, there exist constants c_1, c_2, and c_3 such that

$$\|\mathbf{u}(t)\| < c_1$$
$$\|\Phi(t, t_0)\| < c_2$$
$$\lim_{t\to\infty} \int_{t_0}^{t} \|\Phi(t, t_0)\Phi^{-1}(t_1, t_0)\mathbf{B}(t_1)\|\, dt_1 < c_3$$

Then, we obtain

$$\|\mathbf{x}(t)\| < c_2\|\mathbf{x}(t_0)\| + c_1c_3$$

Hence we see that $\mathbf{x}(t)$ is bounded.

Note that asymptotic stability of the equilibrium state does not necessarily imply the bounded output for a bounded input. For example, consider the following system

$$\dot{x} = -\frac{1}{1+t}x, \qquad x(t_0) = x_0, \qquad t, t_0 > -1$$

which has the solution

$$x(t) = \frac{1+t_0}{1+t}\, x_0$$

The origin of this system is clearly asymptotically stable. Now consider the following equation:

$$\dot{x} = -\frac{1}{1+t}\, x + u, \qquad x(0) = x_0 \tag{8–20}$$

where u is a constant input. Equation (8–20) has the solution

$$x(t) = \frac{1}{1+t}\, x_0 + \frac{t[1+(t/2)]}{1+t}\, u$$

If $u(t) = c_1 =$ nonzero constant, clearly $x(t) \to (c_1/2)t$ as $t \to \infty$, and $x(t)$ of Eq. (8–20) is not bounded.

Similarly, the bounded output for a bounded input does not necessarily imply asymptotic stability of the equilibrium state, $\mathbf{x} = \mathbf{0}$, when the input is removed. For example, consider the following system:

$$\dot{x} = x, \qquad x(0) = x_0 \tag{8–21}$$

which has the solution

$$x(t) = e^t\, x_0$$

Clearly the origin of Eq. (8–21) is unstable. Since, however, for a constant forcing function u the following equation

$$\dot{x} = x + u \tag{8–22}$$

has the solution

$$x(t) = e^t\, x_0 + u\,(e^t - 1) \qquad (t > 0)$$

if we choose $u(t)$ such that $u(t) = -x_0 = c_1$, then $x(t) = -c_1$. The output $x(t)$ of Eq. (8–22) is thus bounded for a bounded input $u(t)$, but the equilibrium state of the free dynamic system is unstable.

8-6 STABILITY ANALYSIS OF NONLINEAR SYSTEMS

In a linear free dynamic system if the equilibrium state is locally asymptotically stable, then it is asymptotically stable in the large. In a nonlinear free dynamic system, however, an equilibrium state can be locally asymptotically stable without being asymptotically stable in the large. Hence implications of asymptotic stability of equilibrium states of linear systems and those of nonlinear systems are quite different.

If we are to examine asymptotic stability of equilibrium states of nonlinear systems, stability analysis of linearized models of nonlinear systems is completely inadequate. We must investigate nonlinear systems without linearization. Several methods based on the second method of Liapunov are available for this purpose (*8–6*). They include Krasovskii's method for testing sufficient conditions for asymptotic stability of nonlinear systems, Schultz-Gibson's variable gradient method for generating Liapunov functions for nonlinear systems, Lur'e's method (*8–7*, *8–8*, *8–9*) applicable to stability analysis of certain nonlinear control systems, Zubov's method (*8–6*) for constructing domain of attraction, and others.

In this section we limit our discussions to Krasovskii's method, Schultz-Gibson's variable gradient method, and Lur'e's method.

Krasovskii's method Consider the nonlinear system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) \tag{8–23}$$

where $\mathbf{x} = n$ vector (state vector)

$\mathbf{f}(\mathbf{x}) = n$ vector whose elements are nonlinear functions of $x_1, x_2, \ldots, x_n$

The Jacobian matrix for the system is

$$\mathbf{F}(\mathbf{x}) = \left[\frac{\partial(f_1, \ldots, f_n)}{\partial(x_1, \ldots, x_n)}\right] = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

In this nonlinear system there may be more than one equilibrium state. It is, however, possible to transfer the equilibrium state under consideration to the origin of the state space by an appropriate transformation of coordinates. We shall, therefore, consider the equilibrium state under consideration to be at the origin.

We shall now present Krasovskii's theorem.

Theorem 8-13 (*Krasovskii's theorem*) Consider the system described by Eq. (8–23). Assume that $\mathbf{f}(\mathbf{0}) = \mathbf{0}$ and that $\mathbf{f}(\mathbf{x})$ is differentiable with respect to x_i, $i = 1, 2, \ldots, n$. Define

$$\hat{\mathbf{F}}(\mathbf{x}) = \mathbf{F}^*(\mathbf{x}) + \mathbf{F}(\mathbf{x})$$

where $\mathbf{F}(\mathbf{x})$ is the Jacobian matrix and $\mathbf{F}^*(\mathbf{x})$ is the conjugate transpose of $\mathbf{F}(\mathbf{x})$. [If $\mathbf{f}(\mathbf{x})$ is real, then $\mathbf{F}(\mathbf{x})$ is real and $\mathbf{F}^*(\mathbf{x})$ can be written as $\mathbf{F}'(\mathbf{x})$. $\hat{\mathbf{F}}(\mathbf{x})$ is clearly Hermitian (symmetric if $\mathbf{F}(\mathbf{x})$ is real).] If $\hat{\mathbf{F}}(\mathbf{x})$ is negative definite,

then the equilibrium state $\mathbf{x} = \mathbf{0}$ is asymptotically stable. A Liapunov function for this system is

$$V(\mathbf{x}) = \mathbf{f}^*(\mathbf{x})\mathbf{f}(\mathbf{x})$$

If, in addition $\mathbf{f}^*(\mathbf{x})\mathbf{f}(\mathbf{x}) \to \infty$ as $\|\mathbf{x}\| \to \infty$, then the equilibrium state is asymptotically stable in the large.

Proof Since $\hat{\mathbf{F}}(\mathbf{x})$ is negative definite, the determinant of $\mathbf{F}$ is nonzero everywhere except at $\mathbf{x} = \mathbf{0}$. (See Prob. A-8-13.) There is no other equilibrium state than $\mathbf{x} = \mathbf{0}$ in the entire state space, or $\mathbf{f}(\mathbf{x}) \neq \mathbf{0}$ for $\mathbf{x} \neq \mathbf{0}$. Since $\mathbf{f}(\mathbf{0}) = \mathbf{0}$, $\mathbf{f}(\mathbf{x}) \neq \mathbf{0}$ for $\mathbf{x} \neq \mathbf{0}$ and $V(\mathbf{x}) = \mathbf{f}^*(\mathbf{x})\mathbf{f}(\mathbf{x})$, $V(\mathbf{x})$ is positive definite.

Noting that

$$\dot{\mathbf{f}}(\mathbf{x}) = \mathbf{F}(\mathbf{x})\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x})\mathbf{f}(\mathbf{x})$$

$\dot{V}$ can be obtained as follows:

$$\begin{aligned}\dot{V}(\mathbf{x}) &= \dot{\mathbf{f}}^*(\mathbf{x})\mathbf{f}(\mathbf{x}) + \mathbf{f}^*(\mathbf{x})\dot{\mathbf{f}}(\mathbf{x}) \\ &= [\mathbf{F}(\mathbf{x})\mathbf{f}(\mathbf{x})]^*\mathbf{f}(\mathbf{x}) + \mathbf{f}^*(\mathbf{x})\mathbf{F}(\mathbf{x})\mathbf{f}(\mathbf{x}) \\ &= \mathbf{f}^*(\mathbf{x})[\mathbf{F}^*(\mathbf{x}) + \mathbf{F}(\mathbf{x})]\mathbf{f}(\mathbf{x}) \\ &= \mathbf{f}^*(\mathbf{x})\hat{\mathbf{F}}(\mathbf{x})\mathbf{f}(\mathbf{x})\end{aligned}$$

Since $\hat{\mathbf{F}}(\mathbf{x})$ is negative definite it follows that $\dot{V}(\mathbf{x})$ is negative definite. Hence $V(\mathbf{x})$ is a Liapunov function. Therefore, the origin is asymptotically stable. If $V(\mathbf{x}) = \mathbf{f}^*(\mathbf{x})\,\mathbf{f}(\mathbf{x})$ tends to infinity as $\|\mathbf{x}\| \to \infty$, then by Theorem 8–2 the equilibrium state is asymptotically stable in the large.

Notice that Krasovskii's theorem differs from usual linearization approaches. It is not limited to small departures from the equilibrium state. Notice also that $V(\mathbf{x})$ and $\dot{V}(\mathbf{x})$ are represented in terms of $\mathbf{f}(\mathbf{x})$ or $\dot{\mathbf{x}}$ rather than in terms of $\mathbf{x}$.

The foregoing theorem for asymptotic stability in the large gives sufficient conditions for nonlinear systems and necessary and sufficient conditions for linear systems. An equilibrium state of a nonlinear system may be stable even if the conditions specified in this theorem are not satisfied. Hence in using Theorem 8–13 be careful not to give erroneous conclusions about the stability of a given nonlinear system.

Example 8-10 (8–2) Consider the second-order system with two nonlinearities:

$$\begin{aligned}\dot{x}_1 &= f_1(x_1) + f_2(x_2) \\ \dot{x}_2 &= x_1 + ax_2\end{aligned} \tag{8-24}$$

We assume that $f_1(0) = f_2(0) = 0$, and that $f_1(x_1)$ and $f_2(x_2)$ are real and differentiable. In addition, assume that

$$[f_1(x_1) + f_2(x_2)]^2 + [x_1 + ax_2]^2 \to \infty \qquad \text{as } \|\mathbf{x}\| \to \infty$$

Determine sufficient conditions for asymptotic stability of the equilibrium state $\mathbf{x} = \mathbf{0}$.

In this system $\mathbf{F}(\mathbf{x})$ is given by

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} f_1'(x_1) & f_2'(x_2) \\ 1 & a \end{bmatrix}$$

where $$f_1'(x_1) = \frac{\partial f_1}{\partial x_1} \qquad \text{and} \qquad f_2'(x_2) = \frac{\partial f_2}{\partial x_2}$$

Then $\hat{\mathbf{F}}(\mathbf{x})$ becomes as follows:

$$\begin{aligned}\hat{\mathbf{F}}(\mathbf{x}) &= \mathbf{F}'(\mathbf{x}) + \mathbf{F}(\mathbf{x}) \\ &= \begin{bmatrix} 2f_1'(x_1) & 1 + f_2'(x_2) \\ 1 + f_2'(x_2) & 2a \end{bmatrix}\end{aligned}$$

By Krasovskii's theorem if $\hat{\mathbf{F}}(\mathbf{x})$ is negative definite, then the equilibrium state $\mathbf{x} = \mathbf{0}$ of the system described by Eq. (8–24) is asymptotically stable in the large. Hence if

(1) $f_1'(x_1) < 0$ for all $x_1 \neq 0$

(2) $4af_1'(x_1) - [1 + f_2'(x_2)]^2 > 0$ for all $x_1 \neq 0$, $x_2 \neq 0$

then the equilibrium state $\mathbf{x} = \mathbf{0}$ is asymptotically stable in the large. (These two conditions are sufficient conditions for asymptotic stability. It is evident that because stability conditions are completely independent of actual forms of the nonlinear functions $f_1(x)$ and $f_2(x)$, these conditions are unduly restrictive.)

Variable gradient method for generating Liapunov functions *(8–10)* For an asymptotically stable system, there are a large or even infinite number of V functions. In fact, it is the existence of this large number of suitable Liapunov functions that gives the second method of Liapunov an advantage over other methods in stability analysis.

The major difficulty in applying the second method of Liapunov to the stability analysis of nonlinear systems is the lack of a straightforward procedure for finding Liapunov functions. Krasovskii's approach assumes V to be a Hermitian form or a quadratic form in $\mathbf{f}(\mathbf{x})$. Such assumption of a Hermitian or quadratic form is unnecessarily restrictive simply because a Hermitian or quadratic V function may not exist for a given system.

One useful systematic approach is the variable gradient method of Schultz and Gibson. This method is based on the fact that if a particular Liapunov function exists which is capable of proving asymptotic stability of a given system, then a unique gradient of this Liapunov function also exists. Through the introduction of a variable gradient and auxiliary equations, this approach to some extent reduces the emphasis on the ingenuity and experience of the investigator.

We shall first present basic theoretical considerations and then explain and illustrate the method by examples.

Consider the system described by the following equation:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t)$$

where $\mathbf{x}$ is a state vector (n vector) and $\mathbf{f}(\mathbf{x}, t)$ is an n vector, elements of which are functions of $x_1, x_2, \ldots, x_n$ and t. We assume the equilibrium state under consideration to be the origin of the state space. Let us denote a tentative Liapunov function as V. In this tentative Liapunov function we assume, for convenience, that V is an explicit function of $x_1, x_2, \ldots, x_n$, but not an explicit function of t. Then

$$\dot{V} = \frac{\partial V}{\partial x_1}\dot{x}_1 + \frac{\partial V}{\partial x_2}\dot{x}_2 + \cdots + \frac{\partial V}{\partial x_n}\dot{x}_n$$

$\dot{V}$ can be determined from ∇V, the gradient of V, as follows:

$$\dot{V} = (\nabla V)'\dot{\mathbf{x}}$$

where $(\nabla V)'$ is the transpose of ∇V. ∇V is given by

$$\nabla V = \begin{bmatrix} \dfrac{\partial V}{\partial x_1} \\ \vdots \\ \dfrac{\partial V}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla V_1 \\ \vdots \\ \nabla V_n \end{bmatrix}$$

V is obtained as line integral of ∇V as

$$V = \int_0^{\mathbf{x}} (\nabla V)' d\mathbf{x} \tag{8-25}$$

The upper limit here is not meant to imply that V is a vector quantity, but rather that the integral is a line integral to an arbitrary point in the state space located at $(x_1, x_2, \ldots, x_n)$. This integral can be made independent of the path of integration. The simplest path of integration is indicated by the expanded form of Eq. (8–25) to be

$$\begin{aligned} V = &\int_0^{x_1,\ (x_2 = x_3 = \cdots = x_n = 0)} \nabla V_1 dx_1 \\ &+ \int_0^{x_2,\ (x_1 = x_1,\ x_3 = x_4 = \cdots = x_n = 0)} \nabla V_2 dx_2 \\ &+ \cdots \\ &+ \int_0^{x_n,\ (x_1 = x_1,\ x_2 = x_2,\ \ldots,\ x_{n-1} = x_{n-1})} \nabla V_n dx_n \end{aligned} \tag{8-26}$$

where ∇V_i is the component of ∇V in the x_i direction.

For a scalar function V to be obtained uniquely from a line integral of a vector function ∇V, the following matrix $\mathbf{F}$ formed by $\partial \nabla V_i / \partial x_j$:

$$\mathbf{F} = \begin{bmatrix} \dfrac{\partial \nabla V_1}{\partial x_1} & \dfrac{\partial \nabla V_1}{\partial x_2} & \cdots & \dfrac{\partial \nabla V_1}{\partial x_n} \\ \dfrac{\partial \nabla V_2}{\partial x_1} & \dfrac{\partial \nabla V_2}{\partial x_2} & \cdots & \dfrac{\partial \nabla V_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial \nabla V_n}{\partial x_1} & \dfrac{\partial \nabla V_n}{\partial x_2} & \cdots & \dfrac{\partial \nabla V_n}{\partial x_n} \end{bmatrix}$$

must be symmetric. In the case of $n = 3$ we have

$$\frac{\partial \nabla V_2}{\partial x_1} = \frac{\partial \nabla V_1}{\partial x_2}, \qquad \frac{\partial \nabla V_3}{\partial x_1} = \frac{\partial \nabla V_1}{\partial x_3}, \qquad \frac{\partial \nabla V_3}{\partial x_2} = \frac{\partial \nabla V_2}{\partial x_3}$$

The condition on the matrix $\mathbf{F}$ is thus a generalized curl requirement for the n-dimensional case. The problem of determining a V function which satisfies Liapunov's theorem is then transformed into the problem of finding a ∇V such that the n-dimensional curl of ∇V equals zero. Further, the V and $\dot{V}$ determined from ∇V must be sufficient to prove stability; that is, they must satisfy Liapunov's theorem. We first set ∇V is equal to an arbitrary column vector

$$\nabla V = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n \end{bmatrix} \tag{8-27}$$

The a_{ij} are completely undetermined quantities. The a_{ij} may be constant or functions of time t and/or functions of state variables. It is convenient, however, to choose a_{nn} as a constant or as a function of time t. Several of the a_{ij} may be chosen to be zero or they are obvious from constraints on $\dot{V}$ imposed by the investigator, or they are determined from the curl equations.

If the equilibrium state $\mathbf{x} = \mathbf{0}$ of a nonlinear system is asymptotically stable, then we may be able to obtain a Liapunov function by the method just described by following the following procedures:

1. Assume ∇V of the form of Eq. (8–27).
2. From ∇V, determine $\dot{V}$.
3. Constrain $\dot{V}$ to be negative definite or at least negative semidefinite.
4. Use the $n(n-1)/2$ curl equations implied by the statement that $\mathbf{F}$ must be symmetric to determine the remaining unknown coefficients in ∇V.
5. Recheck $\dot{V}$, as the addition of terms required as a result of step 4 may alter $\dot{V}$.

6. Determine V by Eq. (8–26).
7. Examine the region of asymptotic stability.

Note that failure to obtain a suitable Liapunov function by this method does not imply instability of the equilibrium state. (In fact no conclusions about stability or instability can be made.) Two examples are given to illustrate the preceding procedure.

Example 8-11 Consider the following system:

$$\dot{x}_1 = -\frac{6x_1}{(1+x_1^2)^2} + 2x_2$$

$$\dot{x}_2 = -2x_1 - \frac{2x_2}{(1+x_1^2)^2}$$

Determine the stability of the equilibrium state $\mathbf{x} = \mathbf{0}$ of the system.

Let us attempt to find a suitable Liapunov function by the variable gradient method. Let

$$\nabla V = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + 2x_2 \end{bmatrix}$$

Then $\dot{V}$ can be obtained as

$$\dot{V} = (\nabla V)'\dot{\mathbf{x}} = -\frac{6a_{11}x_1^2}{(1+x_1^2)^2} + 2a_{12}x_2^2 + \left(2a_{11} - \frac{6a_{12}}{(1+x_1^2)^2}\right)x_1x_2 - 2a_{21}x_1^2 - \frac{4x_2^2}{(1+x_1^2)^2} - \left(4 + \frac{2a_{21}}{(1+x_1^2)^2}\right)x_1x_2$$

Here, there are a number of ways in which $\dot{V}$ might be constrained in order to prove stability. For the purpose of proving stability, $\dot{V}$ must be at least negative semidefinite. This can be accomplished by choosing, for example, $a_{12} = a_{21} = 0$ and $a_{11} = 2$. With these values of a_{11}, a_{12}, and a_{21} we obtain

$$\dot{V} = -\frac{12x_1^2}{(1+x_1^2)^2} - \frac{4x_2^2}{(1+x_1^2)^2}$$

from which we see that $\dot{V}$ is negative definite. The function V is then obtained from

$$V = \int_0^{x_1,\,(x_2=0)} 2x_1\,dx_1 + \int_0^{x_2,\,(x_1=x_1)} 2x_2\,dx_2 = x_1^2 + x_2^2$$

We see that V is positive definite. Hence V is a Liapunov function. Since $V \to \infty$ as $\|\mathbf{x}\| \to \infty$ the equilibrium state $\mathbf{x} = \mathbf{0}$ is asymptotically stable in the large.

Example 8-12 (8-11) Examine the stability of the equilibrium state $\mathbf{x} = \mathbf{0}$ of the following system:

$$\dot{x}_1 = x_2$$

$$\dot{x}_2 = x_3$$

$$\dot{x}_3 = -(x_1 + bx_2)^3 - cx_3$$

Let us attempt to construct a Liapunov function by the variable gradient method. Let the gradient of V be defined by

$$\nabla V = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\ a_{31}x_1 + a_{32}x_2 + x_3 \end{bmatrix}$$

then the derivative of V becomes

$$\begin{aligned}\dot{V} &= (\nabla V)'\dot{\mathbf{x}} \\ &= (a_{11}x_1 + a_{12}x_2 + a_{13}x_3)x_2 \\ &\quad + (a_{21}x_1 + a_{22}x_2 + a_{23}x_3)x_3 \\ &\quad - (a_{31}x_1 + a_{32}x_2 + x_3)cx_3 \\ &\quad - (a_{31}x_1 + a_{32}x_2 + x_3)(x_1 + bx_2)^3\end{aligned}$$

$\dot{V}$ might be constrained in a number of ways in order to prove stability. By looking at the equation for $\dot{V}$ we first choose

$$a_{23} = c, \qquad a_{32} = c, \qquad a_{22} = c^2, \qquad a_{31} = 0, \qquad a_{13} = 0$$

Then ∇V is simplified as

$$\nabla V = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + c^2x_2 + cx_3 \\ cx_2 + x_3 \end{bmatrix}$$

The values of the a's must be such that the following curl equations are satisfied:

$$\frac{\partial \nabla V_1}{\partial x_2} = \frac{\partial \nabla V_2}{\partial x_1}, \qquad \frac{\partial \nabla V_1}{\partial x_3} = \frac{\partial \nabla V_3}{\partial x_1}, \qquad \frac{\partial \nabla V_2}{\partial x_3} = \frac{\partial \nabla V_3}{\partial x_2} \tag{8-28}$$

or

$$\begin{aligned}&\frac{\partial(a_{11}x_1)}{\partial x_2} + \frac{\partial(a_{12}x_2)}{\partial x_2} = \frac{\partial(a_{21}x_1)}{\partial x_1} \\ &\frac{\partial(a_{11}x_1)}{\partial x_3} + \frac{\partial(a_{12}x_2)}{\partial x_3} = 0 \\ &\frac{\partial(a_{21}x_1)}{\partial x_3} + c = c\end{aligned} \tag{8-29}$$

From Eq. (8–29) it is clear that a_{21} does not involve x_3. As a trial, choose

$$a_{21} = \frac{1}{x_1}(x_1 + bx_2)^3$$

Then

$$\frac{\partial(a_{21}x_1)}{\partial x_1} = 3(x_1 + bx_2)^2$$

If we choose

$$a_{12} = \frac{1}{bx_2}(x_1 + bx_2)^3$$

then,

$$\frac{\partial(a_{12}x_2)}{\partial x_2} = 3(x_1 + bx_2)^2$$

If a_{11} is a function only of x_1, then such choices of a_{21} and a_{12} would satisfy Eq. (8–29). Hence we choose ∇V as follows:

$$\nabla V = \begin{bmatrix} a_{11}(x_1)x_1 + \frac{1}{b}(x_1 + bx_2)^3 \\ (x_1 + bx_2)^3 + c^2x_2 + cx_3 \\ cx_2 + x_3 \end{bmatrix} \tag{8–30}$$

then the derivative of V becomes

$$\begin{aligned} \dot{V} &= (\nabla V)'\dot{\mathbf{x}} \\ &= \left[a_{11}(x_1)x_1 + \frac{1}{b}(x_1 + bx_2)^3\right]x_2 \\ &\quad + [(x_1 + bx_2)^3 + c^2x_2 + cx_3]x_3 \\ &\quad + (cx_2 + x_3)[-(x_1 + bx_2)^3 - cx_3] \\ &= (b^2 - cb^3)x_2^4 + 3(1 - bc)x_1^2x_2^2 \\ &\quad + 3(b - b^2c)x_1x_2^3 + \left(\frac{1}{b} - c\right)x_2x_1^3 \\ &\quad + a_{11}(x_1)x_1x_2 \end{aligned}$$

If we choose

$$a_{11}(x_1) = \left(c - \frac{1}{b}\right)x_1^2$$

then

$$\dot{V} = x_2^2(1 - bc)(b^2x_2^2 + 3x_1^2 + 3bx_1x_2)$$

Note that the quantity $(3x_1^2 + 3bx_1x_2 + b^2x_2^2)$ is positive definite, since

$$3 > 0, \qquad \begin{vmatrix} 3 & 1.5b \\ 1.5b & b^2 \end{vmatrix} > 0$$

Hence for the negative definite condition on $\dot{V}$, we have

$$bc > 1 \tag{8–31}$$

∇V defined by Eq. (8–30) can now be written as

$$\nabla V = \begin{bmatrix} \left(c - \frac{1}{b}\right)x_1^3 + \frac{1}{b}(x_1 + bx_2)^3 \\ (x_1 + bx_2)^3 + c^2x_2 + cx_3 \\ cx_2 + x_3 \end{bmatrix}$$

V is found by the following line integration

$$\begin{aligned} V &= \int_0^{\mathbf{x}} (\nabla V)'\, d\mathbf{x} \\ &= \int_0^{x_1,\ (x_2 = x_3 = 0)} cx_1^3\, dx_1 + \int_0^{x_2,\ (x_1 = x_1,\ x_3 = 0)} [(x_1 + bx_2)^3 + c^2x_2]\, dx_2 \\ &\quad + \int_0^{x_3,\ (x_1 = x_1,\ x_2 = x_2)} (cx_2 + x_3)dx_3 \\ &= \frac{c}{4}x_1^4 + \frac{1}{4b}(x_1 + bx_2)^4 + \frac{1}{2}(cx_2 + x_3)^2 \end{aligned}$$

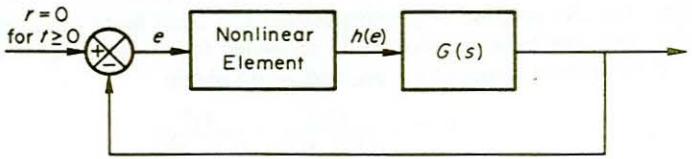
Clearly, V is positive definite if

$$b > 0, \qquad c > 0$$

Note that $V(\mathbf{x}) \to \infty$ as $\|\mathbf{x}\| \to \infty$.

Since from inequality (8–31), $\dot{V}$ is negative definite if $bc > 1$, we conclude that V is a Liapunov function under the condition that $b > 0$, $c > 0$, $bc > 1$. Therefore, the equilibrium state $\mathbf{x} = \mathbf{0}$ is asymptotically stable if $b > 0$, $c > 0$, and $bc > 1$. (Note that these are sufficient conditions for asymptotic stability.)

Lur'e's method We present next a stability analysis of nonlinear systems by use of the first canonic form of Lur'e. A. I. Lur'e has studied, by means of the second method of Liapunov, a class of systems with one nonlinear element. An advantage of this approach is that V functions have been developed for nonlinear systems in the canonic form. Major disadvantages of this approach are that it requires a rather tedious job of finding a solution to a set of algebraic equations, and that this approach can be applied to only a rather limited class of nonlinear systems. In order to save tedious computations for a solution to algebraic equations, various



$$G(s) = \frac{-a_1}{s - \lambda_1} + \frac{-a_2}{s - \lambda_2} + \frac{-a_3}{s - \lambda_3} + \cdots$$

λ_j : Simple poles in the left half plane excluding the origin

Figure 8-6 Nonlinear control system.

simplified stability criteria have been developed in the literature (*8–12*, pp. 324–26) which would facilitate the stability analysis of nonlinear systems in the canonic form of Lur'e.

Consider the control system shown in Fig. 8–6. It is assumed that $G(s)$ involves only simple nonzero poles in the left half plane. The input-output characteristic curve of the nonlinear element is assumed to be a continuous, single-valued function. The output of the nonlinear element is $h(e)$, where e is the input to the nonlinear element. It is assumed that $h(0) = 0$. The nonlinear element may consist of a controller and other elements. Such a system having only one nonlinear element can be described in the first canonic form. In the system shown in Fig. 8–6, it is asumed that the input to the system, $r(t)$, is removed at time $t = 0$ and $r(t) = 0$ for all $t > 0$.

The first canonic form of Lur'e The nonlinear control system shown in Fig. 8–6 may be described in the following first canonic form of Lur'e:

$$\begin{aligned} \dot{x}_i &= \lambda_i x_i + h(e) \qquad (i = 1, 2, \ldots, n) \\ e &= \sum_{i=1}^{n} \alpha_i x_i \end{aligned} \tag{8–32}$$

where x_i = canonic variable
e = input to the nonlinear element
$h(e)$ = output of the nonlinear element [the input to $G(s)$]
λ_i = ith pole of $G(s)$ (λ_i can be either real or complex)
α_i = negative of the residue at pole $s = \lambda_i$

Notice that $\dot{e}$ can be given by

$$\dot{e} = \sum_{i=1}^{n} \alpha_i \lambda_i x_i - r h(e)$$

where

$$r = -\sum_{i=1}^{n} \alpha_i$$

For real λ's the corresponding canonic variables become real. For complex conjugate λ's, the corresponding canonic variables are complex conjugate pairs.

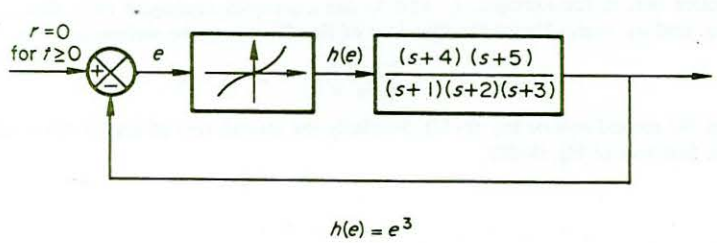


Figure 8-7 Nonlinear control system.

As an example consider the system shown in Fig. 8–7. Equations for this system in the first canonic form can be obtained as follows: Since $G(s)$ can be written as

$$\begin{aligned} G(s) &= \frac{(s+4)(s+5)}{(s+1)(s+2)(s+3)} \\ &= \frac{6}{s+1} - \frac{6}{s+2} + \frac{1}{s+3} \end{aligned}$$

we obtain

$$\begin{aligned} &\lambda_1 = -1, \quad \lambda_2 = -2, \quad \lambda_3 = -3 \\ &\alpha_1 = -6, \quad \alpha_2 = 6, \quad \alpha_3 = -1 \\ &r = -(\alpha_1 + \alpha_2 + \alpha_3) = 1 \end{aligned}$$

Hence the system equations in the first canonic form is given by

$$\begin{aligned} \dot{x}_1 &= -x_1 + h(e) \\ \dot{x}_2 &= -2x_2 + h(e) \\ \dot{x}_3 &= -3x_3 + h(e) \end{aligned} \tag{8–33}$$

where $e = -6x_1 + 6x_2 - x_3, \qquad h(e) = e^3$

By direct computation

$$\dot{e} = 6x_1 - 12x_2 + 3x_3 - h(e)$$

Once the numerical values for the coefficients λ_i, α_i, and r have been calculated it is possible to obtain stability information of this nonlinear control system.

It is convenient to write Eq. (8–32) in the form of the vector matrix differential equation as follows:

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + h(e)\mathbf{y} \\ e &= \mathbf{z}^*\mathbf{x} \end{aligned} \tag{8–34}$$

where $\mathbf{x} = n$ vector
$\mathbf{A} = n \times n$ diagonal matrix whose diagonal elements are $\lambda_1, \lambda_2, \ldots, \lambda_n$ where the λ_i are negative real quantities and/or complex conjugate pairs having negative real parts

$$\mathbf{y} = n \text{ vector} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

e = scalar, the input to the nonlinear element
$h(e)$ = scalar, the output of the nonlinear element

$$\mathbf{z} = n \text{ vector} = \begin{bmatrix} \bar{\alpha}_1 \\ \bar{\alpha}_2 \\ \vdots \\ \bar{\alpha}_n \end{bmatrix}$$

Then $\dot{e}$ can be given by

$$\dot{e} = \sum_{i=1}^{n} \alpha_i \lambda_i x_i - rh(e) = \mathbf{z}^* \mathbf{A} \mathbf{x} - rh(e)$$

$$= \sum_{i=1}^{n} \bar{\alpha}_i \bar{\lambda}_i \bar{x}_i - rh(e) = \mathbf{x}^* \mathbf{A}^* \mathbf{z} - rh(e) \qquad (8\text{–}35)$$

Sufficient conditions for asymptotic stability of the origin of the system given by Eq. (8–34) can be stated as follows:

Theorem 8-14 (*Lur'e's theorem*) The origin of the system given by Eq. (8–34) is asymptotically stable in the large if

(1) $r = -\sum_{i=1}^{n} \alpha_i > 0$

(2) $h(e)$ is continuous for all values of e, $eh(e) \geqslant 0$ for all values of $e \neq 0$, $\int_0^e h(e)\, de \to \infty$ as $|e| \to \infty$ and $h(0) = 0$

(3) there exists at least one set of $a_1, a_2, \ldots, a_n$ which satisfies the following matrix equation:

$$\begin{bmatrix} 2\sqrt{r}\, a_1 - 2\sum_{i=1}^{n} \frac{a_1 a_i}{\lambda_1 + \lambda_i} + \alpha_1 \lambda_1 \\ 2\sqrt{r}\, a_2 - 2\sum_{i=1}^{n} \frac{a_2 a_i}{\lambda_2 + \lambda_i} + \alpha_2 \lambda_2 \\ \vdots \\ 2\sqrt{r}\, a_n - 2\sum_{i=1}^{n} \frac{a_n a_i}{\lambda_n + \lambda_i} + \alpha_n \lambda_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad (8\text{–}36)$$

where a_j is real if λ_j is real, and a_k and a_{k+1} are complex conjugates if λ_k and λ_{k+1} are complex conjugates.

If the first condition $r > 0$ is modified to $r \geqslant 0$, then the origin is asymptotically stable in the large if $\dot{V}$ does not vanish identically in $t \geqslant 0$ for any $\mathbf{x}_0 \neq \mathbf{0}$.

Proof If there exists at least one solution to Eq. (8–36) satisfying the condition that a_j is real if λ_j is real and a_k and a_{k+1} are complex conjugates if λ_k and λ_{k+1} are complex conjugates, then Eq. (8–36) can be rewritten as follows:

$$\begin{bmatrix} 2\sqrt{r}\, \bar{a}_1 - 2\sum_{i=1}^{n} \frac{\bar{a}_1 a_i}{\bar{\lambda}_1 + \lambda_i} + \bar{\alpha}_1 \bar{\lambda}_1 \\ 2\sqrt{r}\, \bar{a}_2 - 2\sum_{i=1}^{n} \frac{\bar{a}_2 a_i}{\bar{\lambda}_2 + \lambda_i} + \bar{\alpha}_2 \bar{\lambda}_2 \\ \vdots \\ 2\sqrt{r}\, \bar{a}_n - 2\sum_{i=1}^{n} \frac{\bar{a}_n a_i}{\bar{\lambda}_n + \lambda_i} + \bar{\alpha}_n \bar{\lambda}_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad (8\text{–}37)^1$$

Equation (8–37) can be simplified as follows:

$$2\sqrt{r}\, \mathbf{v} + 2\mathbf{S}\mathbf{y} + \mathbf{A}^* \mathbf{z} = \mathbf{0}$$

where

$$\mathbf{v} = \begin{bmatrix} \bar{a}_1 \\ \bar{a}_2 \\ \vdots \\ \bar{a}_n \end{bmatrix}, \qquad \mathbf{S} = \begin{bmatrix} -\frac{\bar{a}_1 a_1}{\bar{\lambda}_1 + \lambda_1} & -\frac{\bar{a}_1 a_2}{\bar{\lambda}_1 + \lambda_2} & \cdots & -\frac{\bar{a}_1 a_n}{\bar{\lambda}_1 + \lambda_n} \\ -\frac{\bar{a}_2 a_1}{\bar{\lambda}_2 + \lambda_1} & -\frac{\bar{a}_2 a_2}{\bar{\lambda}_2 + \lambda_2} & \cdots & -\frac{\bar{a}_2 a_n}{\bar{\lambda}_2 + \lambda_n} \\ \vdots & \vdots & & \vdots \\ -\frac{\bar{a}_n a_1}{\bar{\lambda}_n + \lambda_1} & -\frac{\bar{a}_n a_2}{\bar{\lambda}_n + \lambda_2} & \cdots & -\frac{\bar{a}_n a_n}{\bar{\lambda}_n + \lambda_n} \end{bmatrix}$$

Hence we obtain

$$2\sqrt{r}\, \mathbf{v}^* + 2\mathbf{y}^* \mathbf{S} + \mathbf{z}^* \mathbf{A} = \mathbf{0} \qquad (8\text{–}38)$$

Let us choose the following V function as a tentative Liapunov function:

$$V(\mathbf{x}, e) = \mathbf{x}^* \mathbf{S} \mathbf{x} + \int_0^e h(e)\, de \qquad (8\text{–}39)$$

It can be shown that this V function is positive definite. (Refer to Prob. A-8-27.) V possesses continuous first partial derivatives. $\dot{V}$ can be obtained by use of Eq. (8–35) as

$$\dot{V} = \dot{\mathbf{x}}^* \mathbf{S} \mathbf{x} + \mathbf{x}^* \mathbf{S} \dot{\mathbf{x}} + h(e)\dot{e}$$

$$= \mathbf{x}^* (\mathbf{A}^* \mathbf{S} + \mathbf{S}\mathbf{A}) \mathbf{x} + h(e)(\mathbf{y}^* \mathbf{S} \mathbf{x} + \mathbf{x}^* \mathbf{S} \mathbf{y} + \mathbf{x}^* \mathbf{A}^* \mathbf{z}) - rh^2(e)$$

Noting that

$$\mathbf{y}^* \mathbf{S} \mathbf{x} = \mathbf{x}^* \mathbf{S} \mathbf{y} \qquad \text{and} \qquad \mathbf{x}^* \mathbf{A}^* \mathbf{z} = \mathbf{z}^* \mathbf{A} \mathbf{x}$$

[1] Notice that if, for example, λ_1 and λ_2 are a complex conjugate pair, then $\lambda_1 = \bar{\lambda}_2$, $a_1 = \bar{a}_2$, and $\alpha_1 = \bar{\alpha}_2$. Hence the first row of Eq. (8–36) can be written as

$$2\sqrt{r}\, \bar{a}_2 - 2\sum_{i=1}^{n} \frac{\bar{a}_2 a_i}{\bar{\lambda}_2 \times \lambda_i} + \bar{\alpha}_2 \bar{\lambda}_2 = 0$$

which is the second row of Eq. (8–37). Similarly the second row of Eq. (8–36) is identical with the first row of Eq. (8–37).

$\dot{V}$ becomes

$$\dot{V} = \mathbf{x}^*(\mathbf{A}^*\mathbf{S} + \mathbf{S}\mathbf{A})\mathbf{x} + h(e)(2\mathbf{y}^*\mathbf{S}\mathbf{x} + \mathbf{z}^*\mathbf{A}\mathbf{x}) - rh^2(e)$$

Since

$$\mathbf{A}^*\mathbf{S} + \mathbf{S}\mathbf{A} = -\begin{bmatrix} \bar{a}_1 a_1 & \bar{a}_1 a_2 & \cdots & \bar{a}_1 a_n \\ \bar{a}_2 a_1 & \bar{a}_2 a_2 & \cdots & \bar{a}_2 a_n \\ \vdots & \vdots & & \vdots \\ \bar{a}_n a_1 & \bar{a}_n a_2 & \cdots & \bar{a}_n a_n \end{bmatrix}$$

$$= \text{negative semidefinite}^1$$

we obtain

$$\mathbf{x}^*(\mathbf{A}^*\mathbf{S} + \mathbf{S}\mathbf{A})\mathbf{x} = -\mathbf{x}^*\mathbf{v}\mathbf{v}^*\mathbf{x}$$

Hence $\dot{V}$ becomes

$$\begin{aligned} \dot{V} &= -\mathbf{x}^*\mathbf{v}\mathbf{v}^*\mathbf{x} - rh^2(e) + h(e)(2\mathbf{y}^*\mathbf{S}\mathbf{x} + \mathbf{z}^*\mathbf{A}\mathbf{x}) \\ &= -[\mathbf{v}^*\mathbf{x} + \sqrt{r}\, h(e)]^*[\mathbf{v}^*\mathbf{x} + \sqrt{r}\, h(e)] \\ &\quad + h(e)(\sqrt{r}\,\mathbf{v}^*\mathbf{x} + \sqrt{r}\,\mathbf{x}^*\mathbf{v} + 2\mathbf{y}^*\mathbf{S}\mathbf{x} + \mathbf{z}^*\mathbf{A}\mathbf{x}) \end{aligned}$$

Noting that

$$\mathbf{v}^*\mathbf{x} = \mathbf{x}^*\mathbf{v}$$

we have

$$\begin{aligned} \dot{V} &= -[\mathbf{v}^*\mathbf{x} + \sqrt{r}\, h(e)]^*[\mathbf{v}^*\mathbf{x} + \sqrt{r}\, h(e)] \\ &\quad + h(e)(2\sqrt{r}\,\mathbf{v}^* + 2\mathbf{y}^*\mathbf{S} + \mathbf{z}^*\mathbf{A})\mathbf{x} \end{aligned}$$

Using Eq. (8–38), $\dot{V}$ is simplified into

$$\dot{V} = -[\mathbf{v}^*\mathbf{x} + \sqrt{r}\, h(e)]^*[\mathbf{v}^*\mathbf{x} + \sqrt{r}\, h(e)] \tag{8–40}$$

From Eq. (8–40) we see that if $r > 0$, $\dot{V}$ is negative semidefinite but does not vanish identically on any solution curve. Hence the assumed V function is a Liapunov function. Moreover, $V(\mathbf{x}, e) \to \infty$ as $\|\mathbf{x}\| + |e| \to \infty$. Hence if $r > 0$ then by Theorem 8–5 the origin is asymptotically stable in the large. If $r = 0$ and $\dot{V}$ does not vanish identically in $t \geqslant 0$ for any $\mathbf{x}_0 \neq \mathbf{0}$, then by Theorem 8–5 the origin is asymptotically stable in the large. This completes the proof.

Notice that $G(s)$ cannot have poles at the origin. The characteristic of the nonlinear element can, however, be quite general within the restrictions stated in the theorem. Notice also that this theorem gives only sufficient conditions for asymptotic stability of the origin of nonlinear systems defined by Eq. (8–34).

[1] See Prob. A-8-27.

Example 8-13 Consider the system shown in Fig. 8–7. The system equations in the first canonic form are given by Eq. (8–33), and written thus

$$\begin{aligned} \dot{x}_1 &= -x_1 + h(e) \\ \dot{x}_2 &= -2x_2 + h(e) \\ \dot{x}_3 &= -3x_3 + h(e) \end{aligned}$$

where $e = -6x_1 + 6x_2 - x_3, \qquad h(e) = e^3$

In this system the equilibrium state is the origin and

$$\begin{aligned} \lambda_1 &= -1, \quad \lambda_2 = -2, \quad \lambda_3 = -3 \\ \alpha_1 &= -6, \quad \alpha_2 = 6, \quad \alpha_3 = -1 \\ r &= -(\alpha_1 + \alpha_2 + \alpha_3) = 1 \end{aligned}$$

$$eh(e) = e^4 \geqslant 0$$

$$\int_0^e h(e)\, de \to \infty \qquad \text{as } |e| \to \infty$$

Equation (8–36) becomes

$$\begin{bmatrix} 2a_1 - 2\left(-\frac{a_1^2}{2} - \frac{a_1 a_2}{3} - \frac{a_1 a_3}{4}\right) + 6 \\ 2a_2 - 2\left(-\frac{a_2 a_1}{3} - \frac{a_2^2}{4} - \frac{a_2 a_3}{5}\right) - 12 \\ 2a_3 - 2\left(-\frac{a_3 a_1}{4} - \frac{a_3 a_2}{5} - \frac{a_3^2}{6}\right) + 3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{8–41}$$

Let us check if there exists at least one solution to Eq. (8–41). Equation (8–41) can be simplified as follows:

$$\begin{aligned} a_1\left(1 + \frac{a_1}{2} + \frac{a_2}{3} + \frac{a_3}{4}\right) &= -3 \\ a_2\left(1 + \frac{a_1}{3} + \frac{a_2}{4} + \frac{a_3}{5}\right) &= 6 \\ a_3\left(1 + \frac{a_1}{4} + \frac{a_2}{5} + \frac{a_3}{6}\right) &= -1.5 \end{aligned} \tag{8–42}$$

Notice that obtaining a solution to Eq. (8–42) is rather time-consuming. [For this reason, various simplified versions of Eq. (8–36) are used for the determination of asymptotic stability.] After some computations we find that the following set of a's will satisfy Eq. (8–42):

$$a_1 = -3.45, \qquad a_2 = 5.83, \qquad a_3 = -1.40$$

This set of a's satisfies the condition that a_j must be real if λ_j is real. Hence, by Theorem 8–14, the origin of the system is asymptotically stable in the large.

Notice again that in any given nonlinear system nonexistence of a set of the a_i that satisfies the condition does not imply any conclusions on stability, since Lur'e's theorem gives a sufficient condition but not a necessary and sufficient condition for asymptotic stability.

8-7 STABILITY ANALYSIS OF DISCRETE-TIME SYSTEMS[1]

In this section we present stability analysis of discrete-time systems. As in the case of continuous-time systems, asymptotic stability is the most important concept of the stability of equilibrium states of discrete-time systems. If the system is described in terms of state variables, asymptotic stability of equilibrium states of the discrete-time system obtained by discretizing the continuous-time system is equivalent to that of the original continuous-time system.

First we present stability theorems for discrete-time systems based on the second method of Liapunov.

Theorem 8-15 Consider the following discrete-time system

$$\mathbf{x}((k+1)T) = \mathbf{f}(\mathbf{x}(kT)) \tag{8-43}$$

where $\mathbf{x} = n$ vector
$\mathbf{f}(\mathbf{x}) = n$ vector with the property that $\mathbf{f}(\mathbf{0}) = \mathbf{0}$
T = sampling period

Suppose that there exists a scalar function $V(\mathbf{x})$ continuous in $\mathbf{x}$, such that

(a) $V(\mathbf{x}) > 0$ for $\mathbf{x} \neq \mathbf{0}$
(b) $\Delta V(\mathbf{x}) < 0$ for $\mathbf{x} \neq \mathbf{0}$

where
$$\Delta V(\mathbf{x}(kT)) = V(\mathbf{x}((k+1)T)) - V(\mathbf{x}(kT)) = V(\mathbf{f}(\mathbf{x}(kT))) - V(\mathbf{x}(kT))$$

(c) $V(\mathbf{0}) = 0$
(d) $V(\mathbf{x}) \to \infty$ as $\|\mathbf{x}\| \to \infty$

Then the equilibrium state $\mathbf{x} = \mathbf{0}$ is asymptotically stable in the large and $V(\mathbf{x})$ is a Liapunov function. Note that in this theorem the condition (b) may be replaced by

(b′) $\Delta V(\mathbf{x}) \leqslant 0$ for all $\mathbf{x}$
(b″) $\Delta V(\mathbf{x})$ does not vanish identically for any solution sequence $\{\mathbf{x}(kT)\}$ satisfying Eq. (8-43)

This means that $\Delta V(\mathbf{x})$ need not be negative definite if it does not vanish identically on any solution sequence of the difference equation. (The maximum invariant set is the origin itself. See Section 8-4.)

[1] This section follows closely R. E. Kalman and J. E. Bertram (*8-2*). Numbers in parentheses refer to the references listed at the end of this chapter.

The proof of this theorem is quite similar to that of Theorem 8-1 and is therefore omitted.

The following two theorems apply to discrete-time linear time-invariant systems.

Theorem 8-16 Consider the discrete-time linear time-invariant system

$$\mathbf{x}((k+1)T) = \mathbf{G}\mathbf{x}(kT)$$

where $\mathbf{x}$ is an n vector and $\mathbf{G}$ is an $n \times n$ constant matrix. A necessary and sufficient condition for the origin $\mathbf{x} = \mathbf{0}$ to be asymptotically stable is that all the eigenvalues of the matrix $\mathbf{G}$ be less than unity in absolute value. (Note that this is equivalent to saying that $\|\mathbf{G}^k\| \to 0$ as $k \to \infty$. We leave the proof of Theorem 8-16 to the reader.)

Theorem 8-17 Consider the discrete-time linear time-invariant system

$$\mathbf{x}((k+1)T) = \mathbf{G}\mathbf{x}(kT) \tag{8-44}$$

where $\mathbf{x}$ is an n vector and $\mathbf{G}$ is an $n \times n$ constant matrix. The equilibrium state $\mathbf{x} = \mathbf{0}$ is asymptotically stable if and only if, given any positive definite Hermitian matrix $\mathbf{Q}$ (or any positive definite real symmetric matrix $\mathbf{Q}$), there exists a positive definite Hermitian matrix $\mathbf{P}$ (or positive definite real symmetric matrix $\mathbf{P}$) such that

$$\mathbf{G}^*\mathbf{P}\mathbf{G} - \mathbf{P} = -\mathbf{Q}$$

Then

$$V(\mathbf{x}) = \mathbf{x}^*\mathbf{P}\mathbf{x}$$

is a Liapunov function for the system of Eq. (8-44), with

$$\Delta V(\mathbf{x}) = -\mathbf{x}^*\mathbf{Q}\mathbf{x}$$

(The proof of this theorem is identical with that of Theorem 8-7, and is therefore omitted.)

Notice that the Hermitian form $\mathbf{x}^*\mathbf{P}\mathbf{x}$ or quadratic form $\mathbf{x}'\mathbf{P}\mathbf{x}$ is positive definite if and only if the absolute values of all the eigenvalues of $\mathbf{G}$ are smaller than unity.

For discrete-time systems a norm $\|\mathbf{x}\|$ may be used as a Liapunov function. This fact is stated as a theorem in the following. Before we present the theorem, however, it is necesary to define a term *contraction*. A function $\mathbf{f}(\mathbf{x})$ is said to be a contraction if

$$\|\mathbf{f}(\mathbf{x})\| < \|\mathbf{x}\|$$

for some set of $\mathbf{x} \neq \mathbf{0}$ and some norm, with the property that $\mathbf{f}(\mathbf{0}) = \mathbf{0}$.

Theorem 8-18 Consider the following discrete-time system

$$\mathbf{x}((k+1)T) = \mathbf{f}(\mathbf{x}(kT)), \qquad \mathbf{f}(\mathbf{0}) = \mathbf{0} \tag{8-45}$$

where $\mathbf{x}$ is an n vector and $\mathbf{f}(\mathbf{x})$ is also an n vector, elements of which are functions of the x_i. Assume that $\mathbf{f}$ is a contraction for all $\mathbf{x}$ and some norm. Then the equilibrium state, the origin, of the system of Eq. (8–45) is asymptotically stable in the large and one of its Liapunov functions is

$$V(\mathbf{x}) = ||\mathbf{x}||$$

Proof If we choose $V(\mathbf{x}) = ||\mathbf{x}||$ which is positive definite, then ΔV can be written as

$$\begin{aligned}\Delta V(\mathbf{x}(kT)) &= V(\mathbf{f}(\mathbf{x}(kT)) - V(\mathbf{x}(kT)) \\ &= ||\mathbf{f}(\mathbf{x})|| - ||\mathbf{x}||\end{aligned}$$

which is negative definite by definition of a contraction. Hence $V(\mathbf{x}) = ||\mathbf{x}||$ is a Liapunov function, and the equilibrium state $\mathbf{x} = \mathbf{0}$ is asymptotically stable in the large.

We next present a theorem for a discrete-time forced dynamic system.

Theorem 8-19 If all solutions of

$$\mathbf{x}((k+1)T) = \mathbf{G}\mathbf{x}(kT)$$

where $\mathbf{x}$ is an n vector and $\mathbf{G}$ is an $n \times n$ constant matrix, tend to zero as $k \to \infty$, then all solutions of the following system:

$$\mathbf{x}((k+1)T) = \mathbf{G}\mathbf{x}(kT) + \mathbf{H}\mathbf{u}(kT) \tag{8-46}$$

where $\mathbf{H}$ is an $n \times r$ constant matrix, are bounded, provided that the forcing vector $\mathbf{u}(kT)$, an r vector, is bounded.

Proof Since $\mathbf{u}(kT)$ is bounded, there exists a positive constant c such that

$$||\mathbf{u}(kT)|| < c \qquad \text{for } k = 0, 1, 2, \ldots$$

The solution of Eq. (8–46) is given by

$$\mathbf{x}(kT) = \Phi(kT)\mathbf{x}(0) + \Phi(kT)\sum_{j=1}^{k} \Phi^{-1}(jT)\mathbf{H}\mathbf{u}((j-1)T)$$

where $$\Phi(kT) = \mathbf{G}^k, \qquad \Phi(kT)\Phi^{-1}(jT) = \mathbf{G}^{k-j}$$

Hence,

$$\begin{aligned}||\mathbf{x}(kT)|| &< ||\mathbf{G}||^k||\mathbf{x}(0)|| + c\sum_{j=1}^{k}||\mathbf{G}||^{k-j}||\mathbf{H}|| \\ &\leqslant ||\mathbf{G}||^k||\mathbf{x}(0)|| + \lim_{k\to\infty} c\sum_{j=1}^{k}||\mathbf{G}||^{k-j}||\mathbf{H}||\end{aligned}$$

Since the origin of the free dynamic system is asymptotically stable, there exist constants m and r, $(0 < r < 1)$, such that

$$0 \leqslant ||\mathbf{G}||^k < mr^k$$

Then

$$\lim_{k\to\infty}\sum_{j=1}^{k}||\mathbf{G}||^{k-j} < \lim_{k\to\infty}\sum_{j=1}^{k} mr^{k-j} = m\frac{r}{1-r}$$

Therefore

$$||\mathbf{x}(kT)|| < m||\mathbf{x}(0)|| + c||\mathbf{H}||m\frac{r}{1-r}$$

We thus have proved that $||\mathbf{x}(kT)||$ is bounded.

8-8 APPLICATIONS OF SECOND METHOD OF LIAPUNOV TO SYSTEM DESIGN[1]

In a conventional way of designing control systems a control system is first designed and then its stability is examined. A different approach first formulates the conditions for stability and then designs the system within those limitations. For a large class of control systems a direct relationship can be shown between Liapunov functions and generalized performance indexes used in synthesis of optimal control systems. If the second method of Liapunov is utilized to form the basis for the design of an optimal controller, then this insures an asymptotically stable system and the system output will be continually driven toward its desired value.

In this section we first focus our attention on the relationship between Liapunov functions and transient responses of the system and then discuss the relationship between Liapunov functions and generalized performance indexes.

The following discussion is based on the fact that Liapunov functions give a measure of the state of the system at any given instant, or they can be regarded as a measure of distance in the state space. For an asymptotically stable system for which a Liapunov function has been obtained, the latter can be used to estimate the rapidity of transient response. Let us assume that the equilibrium state is the origin of the state space and define

$$\eta = -\frac{\dot{V}(\mathbf{x}, t)}{V(\mathbf{x}, t)} \tag{8-47}$$

in some region of the state space excluding the origin. For uniformly asymptotically stable systems, a Liapunov function with negative definite $\dot{V}$ can be found, at least in principle. In the following analysis we, therefore, exclude

[1] This section follows closely R. E. Kalman and J. E. Bertram, (*8–2*).

the case where $\dot{V}(\mathbf{x}, t) = 0$ for $\mathbf{x} \neq \mathbf{0}$. Equation (8–47) gives an indication of how fast the system state approaches its equilibrium state. Integrating both sides of Eq. (8–47) with respect to t

$$V(\mathbf{x}, t) = V(\mathbf{x}_0, t_0) e^{-\int_{t_0}^{t} \eta \, dt}$$

where $V(\mathbf{x}_0, t_0)$ corresponds to the initial value of V starting at $\mathbf{x} = \mathbf{x}_0$ at $t = t_0$. If we write the minimum of the ratio $-\dot{V}(\mathbf{x}, t)/V(\mathbf{x}, t)$ as $\eta_{\min}$, or

$$\eta_{\min} = \min\left[-\frac{\dot{V}(\mathbf{x}, t)}{V(\mathbf{x}, t)}\right] \tag{8–48}$$

then

$$\dot{V}(\mathbf{x}, t) \leqslant -\eta_{\min} V(\mathbf{x}, t)$$

and

$$V(\mathbf{x}, t) \leqslant V(\mathbf{x}_0, t_0) e^{-\eta_{\min}(t - t_0)} \tag{8–49}$$

Interpreting V as the distance from the origin, it is clear that Eq. (8–49) gives an estimate of how fast equilibrium is approached.

From Eq. (8–48) we see that $1/\eta_{\min}$ corresponds to the largest time constant relating to changes in the Liapunov function V. Since $V(\mathbf{x})$ may be expanded in series starting with quadratic terms in $\mathbf{x}$, this time constant $1/\eta_{\min}$ is about half the conventional time constant defined for the system. The value of $\eta_{\min}$ thus can be considered as a kind of figure of merit. Larger values of $\eta_{\min}$ correspond to faster response. Since there may be a number of Liapunov functions for a given system, a number of different values of $\eta_{\min}$ may be obtained. The largest value of $\eta_{\min}$ should be considered to represent the figure of merit. It is, however, usually not known how to construct the Liapunov function which yields the largest value of η. The value of η in a sense gives a measure of "goodness" of the Liapunov function for a given system, since this value can be interpreted as the upper bound on the time constant of the system.

Since a Liapunov function for a given system may be obtained without solving the system differential equation, it may be possible to find this figure of merit for a nonlinear system whose solution cannot be found directly. For a nonlinear system, $\eta_{\min}$ may depend on the state of the system. However, this approach is useful only if the Liapunov function available is of a sufficiently simple form.

Calculation of η for linear time-invariant systems (*8–2*) Consider the following linear time-invariant system:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$$

where $\mathbf{x} = n$ vector (state vector)

$\mathbf{A} = n \times n$ constant matrix

Suppose that the eigenvalues of $\mathbf{A}$ have negative real parts. Then a Liapunov function and its time derivative are given by

$$V(\mathbf{x}) = \mathbf{x}^*\mathbf{P}\mathbf{x} \qquad \text{and} \qquad \dot{V}(\mathbf{x}) = -\mathbf{x}^*\mathbf{Q}\mathbf{x}$$

where $\mathbf{Q} = -[\mathbf{A}^*\mathbf{P} + \mathbf{P}\mathbf{A}]$. A definition of $\eta_{\min}$ equivalent to Eq. (8–48) is then

$$\eta_{\min} = \min_{\mathbf{x}} \{\mathbf{x}^*\mathbf{Q}\mathbf{x};\ \mathbf{x}^*\mathbf{P}\mathbf{x} = 1\}$$

The minimization can be performed by use of the Lagrange multiplier technique. Let μ be the Lagrange multiplier. Then, minimizing $\mathbf{x}^*(\mathbf{Q} - \mu\mathbf{P})\mathbf{x}$ with respect to $\mathbf{x}$ implies that the minimum occurs at a value $\mathbf{x}_{\min}$ of $\mathbf{x}$ such that

$$(\mathbf{Q} - \mu\mathbf{P})\mathbf{x}_{\min} = \mathbf{0} \tag{8–50}$$

Therefore,

$$\mathbf{x}^*_{\min}\mathbf{Q}\mathbf{x}_{\min} = \mu\mathbf{x}^*_{\min}\mathbf{P}\mathbf{x}_{\min} = \mu > 0$$

which is a minimum if μ is. But, by Eq. (8–50), μ is an eigenvalue of the matrix $\mathbf{Q}\mathbf{P}^{-1}$. Hence $\eta_{\min}$ is the minimum eigenvalue $\lambda_{\min}$ of $\mathbf{Q}\mathbf{P}^{-1}$, or

$$\eta_{\min} = \lambda_{\min} \tag{8–51}$$

The Liapunov function or V function can thus be used for the direct synthesis of dynamic systems.

Example 8-14 Consider the second-order system:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

This system corresponds to a second-order system with the damping ratio $\zeta = 0.5$ and the undamped natural frequency $\omega_n = 1$. Let us obtain an estimate of the largest time constant of the system by use of a V function.

Let us assume

$$V = \mathbf{x}'\mathbf{P}\mathbf{x}$$

where

$$\mathbf{A}'\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{I}$$

or noting that $\mathbf{P}$ is symmetric, namely $p_{21} = p_{12}$

$$\begin{bmatrix} 0 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} + \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

from which we obtain

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} \frac{3}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$$

Hence

$$V = \mathbf{x}'\mathbf{P}\mathbf{x} = \frac{1}{2}[(x_1 + x_2)^2 + 2x_1^2 + x_2^2]$$

and

$$\dot{V} = -\mathbf{x}'\mathbf{x} = -(x_1^2 + x_2^2)$$

Thus,

$$\eta = -\frac{\dot{V}}{V} = \frac{2(x_1^2 + x_2^2)}{3x_1^2 + 2x_1x_2 + 2x_2^2}$$

Differentiating η with respect to x_1

$$\frac{\partial \eta}{\partial x_1} = \frac{4x_1^2x_2 - 4x_1x_2^2 - 4x_2^3}{(3x_1^2 + 2x_1x_2 + 2x_2^2)^2}$$

Putting $\partial\eta/\partial x_1 = 0$, we obtain

$$x_1 = 1.618\,x_2 \qquad \text{or} \qquad x_1 = -0.618\,x_2$$

Substituting these two values of x_1 into the expression for η we obtain

$$\eta_{\min} = 0.553, \qquad \eta_{\max} = 1.447$$

and the upper limit on the convergence time is thus

$$\frac{1}{\eta_{\min}} = \frac{1}{0.553} = 1.81 \text{ (sec)}$$

This is an estimate of the convergence time for $V(\mathbf{x})$. Since $V(\mathbf{x})$ involves quadratic terms of x_i, $V(\mathbf{x})$ converges about twice as fast as $x_1(t)$ and thus the upper limit on the time constant for $x_1(t)$ would be

$$\tau = \frac{2}{\eta_{\min}} = \frac{2}{0.553} = 3.62 \text{ (sec)}$$

Comparing this figure to the time constant of the given system

$$T = \frac{1}{\zeta\omega_n} = 2$$

we find that the system time constant is less than the upper bound previously predicted namely,

$$2 < 3.62$$

Thus, we expect that possibly some other $V(\mathbf{x})$ would give a closer estimate of the time constant. Notice that it is always not less than the actual system time constant.

$\eta_{\min}$ and $\eta_{\max}$ can also be obtained from Eq. (8–51). Let us first find the eigenvalues λ_i of $\mathbf{QP}^{-1}$, or solve

$$|\mathbf{QP}^{-1} - \lambda\mathbf{I}| = 0$$

Alternatively, using $\mathbf{I}$ instead of $\mathbf{Q}$ solve for the eigenvalues λ_i of $\mathbf{P}^{-1}$ or solve

$$|\mathbf{I} - \lambda\mathbf{P}| = 0$$

In the present example, solve

$$\begin{vmatrix} 1 - \frac{3}{2}\lambda & -\frac{1}{2}\lambda \\ -\frac{1}{2}\lambda & 1 - \lambda \end{vmatrix} = 0$$

The result is

$$\lambda_1 = 1.447, \qquad \lambda_2 = 0.553$$

Hence

$$\eta_{\min} = 0.553 \qquad \text{and} \qquad \eta_{\max} = 1.447$$

This solution, of course, agrees with the solution obtained by directly computing the maximum and minimum values of $-\dot{V}/V$.

We shall next discuss the relationship between Liapunov functions and generalized performance indexes. Suppose that the system is given by

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$$

where $\mathbf{x} = n$ vector

$\mathbf{A} = n \times n$ constant matrix whose eigenvalues have all negative real parts

and the performance index we wish to minimize is given by

$$\int_{t_0}^{\infty} \mathbf{x}^*(t)\mathbf{Q}\mathbf{x}(t)dt$$

where $\mathbf{Q}$ is a positive definite (or semidefinite) Hermitian matrix [or a positive definite (or semidefinite) real symmetric matrix]. For the present system we can choose a Liapunov function and its time derivative as follows:

$$V(\mathbf{x}) = \mathbf{x}^*\mathbf{P}\mathbf{x} \qquad \text{and} \qquad \dot{V}(\mathbf{x}) = -\mathbf{x}^*\mathbf{Q}\mathbf{x}$$

where the positive definite Hermitian matrix $\mathbf{P}$ (or positive definite real symmetric matrix $\mathbf{P}$) is determined from

$$\mathbf{A}^*\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{Q}$$

Then noting that $\mathbf{x}(\infty) = \mathbf{0}$ and $V(\mathbf{0}) = 0$ the given performance index can be written as

$$\begin{aligned} &\int_{t_0}^{\infty} \mathbf{x}^*(t)\mathbf{Q}\mathbf{x}(t)dt \\ &= -\int_{t_0}^{\infty} \dot{V}(\mathbf{x})dt \\ &= -V(\mathbf{x}(\infty)) + V(\mathbf{x}(t_0)) \\ &= V(\mathbf{x}(t_0)) \\ &= \mathbf{x}^*(t_0)\mathbf{P}\mathbf{x}(t_0) \end{aligned}$$

Thus the given performance index equals the value of the Liapunov function

$V(\mathbf{x}) = \mathbf{x}^*\mathbf{P}\mathbf{x}$ at $t = t_0$. This relationship between Liapunov functions and quadratic or Hermitian performance indexes is quite useful in synthesizing optimal control systems. Since we shall discuss this matter in Chapter 9, here we give a simple example illustrating the usefulness of Liapunov functions in the optimization of control systems.

Example 8-15 (8-2) Consider the second-order system given by

$$\ddot{x}_1 + 2\zeta\dot{x}_1 + x_1 = 0, \qquad \zeta > 0 \tag{8-52}$$

where ζ is an undetermined constant. We wish to minimize the performance index $V(\mathbf{x}, \zeta)$ defined by

$$V(\mathbf{x}, \zeta) = \int_0^\infty \mathbf{x}'(t)\mathbf{Q}\mathbf{x}(t)\,dt$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \qquad \mathbf{Q} = \begin{bmatrix} 1 & 0 \\ 0 & \mu \end{bmatrix}, \qquad \mu \geqslant 0$$

Determine the value of ζ so as to minimize the performance index.

Equation (8-52) can be rewritten as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & -2\zeta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

or

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -1 & -2\zeta \end{bmatrix}$$

Since $\mathbf{Q}$ is a positive definite (or semidefinite) real symmetric matrix we can choose a Liapunov function as follows:

$$V(\mathbf{x}) = \mathbf{x}'\mathbf{P}\mathbf{x} \qquad \text{with} \qquad \dot{V}(\mathbf{x}) = -\mathbf{x}'\mathbf{Q}\mathbf{x}$$

where the positive definite real symmetric matrix $\mathbf{P}$ is determined from

$$\mathbf{A}'\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{Q}$$

Solving for $\mathbf{P}$ we obtain

$$\mathbf{P}(\zeta) = \begin{bmatrix} \zeta + \dfrac{1+\mu}{4\zeta} & \dfrac{1}{2} \\ \dfrac{1}{2} & \dfrac{1+\mu}{4\zeta} \end{bmatrix}$$

Since $V(\mathbf{x}, \zeta)$ is equal to $\mathbf{x}'(0)\mathbf{P}\mathbf{x}(0)$

$$V(\mathbf{x}, \zeta) = \zeta x_1^2(0) + \left[\frac{1+\mu}{4\zeta}\right][x_1^2(0) + x_2^2(0)] + x_1(0)x_2(0)$$

The value of $\zeta(\mu)$ which minimizes V when $x_2(0) = 0$ is obtained as follows:

$$\zeta_{\text{opt}}(\mu) = \frac{\sqrt{1+\mu}}{2}$$

from which

$$\zeta_{\text{opt}}(0) = \tfrac{1}{2}$$

This is a well-known result. This type of optimization (minimizing $\int_0^\infty x_1^2\,dt$) yields a system which is too oscillatory. The integral squared error criterion has been known to yield insufficient damping and hence to be impractical. If, however, μ is chosen to be unity,

$$\zeta_{\text{opt}}(1) = \frac{1}{\sqrt{2}} = 0.707$$

This value of ζ is considered to be quite satisfactory. Hence, we see that

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

or minimizing

$$\int_0^\infty (x_1^2 + x_2^2)\,dt$$

may be considered a better error criterion than

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

or minimizing

$$\int_0^\infty x_1^2\,dt$$

8-9 CONCLUDING COMMENTS

The preceding sections of this chapter have presented a reasonably detailed account of stability analysis of dynamic systems and have illustrated many stability and design problems to which the second method of Liapunov is most useful.

The second method of Liapunov proves to be particularly powerful in the investigation of stability analysis of certain nonlinear systems where conventional criteria cannot give answers.

Stability analysis by the second method of Liapunov is simply a problem of finding a positive definite V function (Liapunov function), whose time derivative, taken in the direction of the solution of a system, is nonpositive. The major problem in applying this method to nonlinear systems is the

construction of Liapunov functions. Methods presently available for constructing Liapunov functions are applicable to relatively simple systems only. Ingenuity is still required to apply this stability criterion to systems of other than simple types. We should note again that Liapunov theorems on stability of nonlinear systems furnish only sufficient conditions. Failure to construct a Liapunov function for a given system does not imply instability.

The second method of Liapunov has another useful feature in addition to stability analysis; it can be used for synthesizing optimal control systems in the time domain.

Liapunov functions have been determined for numerous specific differential equations and for certain differential equations of general types.[1] The accompanying exercise problems and solutions include some examples of the specific systems for which Liapunov functions are known.

EXERCISE PROBLEMS AND SOLUTIONS

Problem A-8-1 The closed loop transfer function of a single-input single-output system is given by

$$\frac{C}{R}(s) = \frac{K}{s(s^2 + s + 1)(s + 2) + K}$$

where K is the gain of the system. By use of Routh's stability criterion determine the range of K for asymptotic stability.

Solution The characteristic equation is obtained as

$$s^4 + 3s^3 + 3s^2 + 2s + K = 0$$

The Routh array of coefficients is

s^4	1	3	K
s^3	3	2	0
s^2	$\frac{7}{3}$	K	
s^1	$2 - \frac{9}{7}K$	0	
s^0	K		

For asymptotic stability, K must be positive and all the coefficients in the first column must be positive. Therefore,

$$\frac{14}{9} > K > 0$$

[1] Readers interested in a more detailed account of the second method of Liapunov should see (*8–6*). Those interested in stability (other than the second method of Liapunov) of differential equation systems should refer, for example, to (*8–13*). Readers interested in

(When $K = 14/9$, the system becomes oscillatory and mathematically the oscillation sustains at a constant amplitude.)

Problem A-8-2 Determine the range of K such that the equilibrium state of a linear system whose characteristic equation is given by

$$\lambda^4 + K\lambda^3 + \lambda^2 + \lambda + 1 = 0$$

is asymptotically stable.

Solution The Routh array of coefficients is

λ^4	1	1	1
λ^3	K	1	0
λ^2	$\frac{K-1}{K}$	1	0
λ^1	$1 - \frac{K^2}{K-1}$	0	
λ^0	1		

For asymptotic stability, the conditions are

$$K > 0$$
$$K - 1 > 0$$
$$1 - \frac{K^2}{K-1} > 0$$

From the first and second conditions K must be greater than one. For $K > 1$ the term $1 - [K^2/(K-1)]$ is always negative, since

$$\frac{K - 1 - K^2}{K-1} = \frac{-1 + K(1-K)}{K-1} < 0$$

Thus, the three conditions cannot be fulfilled simultaneously. Therefore, there is no value of K allowing asymptotic stability of the equilibrium state of the system.

Problem A-8-3 Show that the Routh's stability criterion and Hurwitz stability criterion are equivalent.

Solution If we write Hurwitz determinants in the following triangular form:

$$\Delta_i = \begin{vmatrix} a_{11} & & & * \\ & a_{22} & & \\ & & \ddots & \\ 0 & & & a_{ii} \end{vmatrix} \qquad (i = 1, 2, \ldots, n)$$

then the Hurwitz conditions for asymptotic stability become

$$\Delta_i = a_{11}a_{22}\cdots a_{ii} > 0 \qquad (i = 1, 2, \ldots, n)$$

which are equivalent to the conditions

We shall show that these conditions are equivalent to

$$a_1 > 0, \quad b_1 > 0, \quad c_1 > 0, \quad \dots$$

where $a_1, b_1, c_1, \dots$ are the elements of the first column in the Routh array.

Consider, for example, the following Hurwitz determinant:

$$\Delta_4 = \begin{vmatrix} a_1 & a_3 & a_5 & a_7 \\ a_0 & a_2 & a_4 & a_6 \\ 0 & a_1 & a_3 & a_5 \\ 0 & a_0 & a_2 & a_4 \end{vmatrix}$$

The determinant is unchanged if we subtract from the ith row k times the jth row. By subtracting from the second row a_0/a_1 times the first row, we obtain

$$\Delta_4 = \begin{vmatrix} a_{11} & a_3 & a_5 & a_7 \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & a_1 & a_3 & a_5 \\ 0 & a_0 & a_2 & a_4 \end{vmatrix}$$

where

$$a_{11} = a_1$$

$$a_{22} = a_2 - \frac{a_0}{a_1} a_3$$

$$a_{23} = a_4 - \frac{a_0}{a_1} a_5$$

$$a_{24} = a_6 - \frac{a_0}{a_1} a_7$$

Similarly, subtracting from the fourth row a_0/a_1 times the third row yields

$$\Delta_4 = \begin{vmatrix} a_{11} & a_3 & a_5 & a_7 \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & a_1 & a_3 & a_5 \\ 0 & 0 & \hat{a}_{43} & \hat{a}_{44} \end{vmatrix}$$

where

$$\hat{a}_{43} = a_2 - \frac{a_0}{a_1} a_3$$

$$\hat{a}_{44} = a_4 - \frac{a_0}{a_1} a_5$$

Next, subtracting from the third row a_1/a_{22} times the second row yields

$$\Delta_4 = \begin{vmatrix} a_{11} & a_3 & a_5 & a_7 \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & \hat{a}_{43} & \hat{a}_{44} \end{vmatrix}$$

where

$$a_{33} = a_3 - \frac{a_1}{a_{22}} a_{23}$$

$$a_{34} = a_5 - \frac{a_1}{a_{22}} a_{24}$$

Finally, subtracting from the last row $\hat{a}_{43}/a_{33}$ times the third row yields

$$\Delta_4 = \begin{vmatrix} a_{11} & a_3 & a_5 & a_7 \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & a_{44} \end{vmatrix}$$

where

$$a_{44} = \hat{a}_{44} - \frac{\hat{a}_{43}}{a_{33}} a_{34}$$

From this analysis, we see that

$$\Delta_4 = a_{11} a_{22} a_{33} a_{44}$$

$$\Delta_3 = a_{11} a_{22} a_{33}$$

$$\Delta_2 = a_{11} a_{22}$$

$$\Delta_1 = a_{11}$$

The following Hurwitz conditions for asymptotic stability

$$\Delta_1 > 0, \quad \Delta_2 > 0, \quad \Delta_3 > 0, \quad \Delta_4 > 0, \quad \dots$$

reduce to the conditions

$$a_{11} > 0, \quad a_{22} > 0, \quad a_{33} > 0, \quad a_{44} > 0, \quad \dots$$

Referring to the Routh's array, we see that

$$a_{11} = a_1$$

$$a_{22} = a_2 - \frac{a_0}{a_1} a_3 = b_1$$

$$a_{33} = a_3 - \frac{a_1}{a_{22}} a_{23} = \frac{a_3 b_1 - a_1 b_2}{b_1} = c_1$$

$$a_{44} = \hat{a}_{44} - \frac{\hat{a}_{43}}{a_{33}} a_{34} = \frac{b_2 c_1 - b_1 c_2}{c_1} = d_1$$

Hence the Hurwitz conditions for asymptotic stability become

$$a_1 > 0, \quad b_1 > 0, \quad c_1 > 0, \quad d_1 > 0, \quad \dots$$

Thus, we have demonstrated that Hurwitz conditions for asymptotic stability can be reduced to Routh's conditions for asymptotic stability. The same argument can be extended to Hurwitz determinants of any order and the equivalence of Routh's stability criterion and Hurwitz stability criterion can be established.

Problem A-8-4 (8–5) Show that the first column of the Routh array of

$$\lambda^n + a_1 \lambda^{n-1} + a_2 \lambda^{n-2} + \cdots + a_{n-1} \lambda + a_n = 0$$

is given by

$$1, \quad \Delta_1, \quad \frac{\Delta_2}{\Delta_1}, \quad \frac{\Delta_3}{\Delta_2}, \quad \dots, \quad \frac{\Delta_n}{\Delta_{n-1}}$$

where

$$\Delta_r = \begin{vmatrix} a_1 & 1 & 0 & 0 & \cdots & 0 \\ a_3 & a_2 & a_1 & 1 & \cdots & 0 \\ a_5 & a_4 & a_3 & a_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ a_{2r-1} & \cdot & \cdot & \cdot & \cdots & a_r \end{vmatrix}$$

$$a_k = 0 \quad \text{if } k > n$$

Solution The Routh array of coefficients has the form

$$\begin{array}{cccccc} 1 & a_2 & a_4 & a_6 & \cdots & a_n \\ a_1 & a_3 & a_5 & \cdots & & \\ b_1 & b_2 & b_3 & \cdots & & \\ c_1 & c_2 & \cdot & & & \\ \vdots & \vdots & \vdots & & & \end{array}$$

The first term in the first column of the Routh array is 1. The next term in the first column is a_1 which is equal to Δ_1. The next term is b_1 which is equal to

$$\frac{a_1 a_2 - a_3}{a_1} = \frac{\Delta_2}{\Delta_1}$$

The next term in the first column is c_1 which is equal to

$$\frac{b_1 a_3 - a_1 b_2}{b_1} = \frac{\left[\dfrac{a_1 a_2 - a_3}{a_1}\right] a_3 - a_1 \left[\dfrac{a_1 a_4 - a_5}{a_1}\right]}{\left[\dfrac{a_1 a_2 - a_3}{a_1}\right]}$$

$$= \frac{a_1 a_2 a_3 - a_3^2 - a_1^2 a_4 + a_1 a_5}{a_1 a_2 - a_3}$$

$$= \frac{\Delta_3}{\Delta_2}$$

In a similar manner the remaining terms in the first column of the Routh array can be found.

The Routh array has the property that the last nonzero terms of any columns are the same, that is, if the array is given by

$$\begin{array}{cccc} a_0 & a_2 & a_4 & a_6 \\ a_1 & a_3 & a_5 & a_7 \\ b_1 & b_2 & b_3 & \\ c_1 & c_2 & c_3 & \\ d_1 & d_2 & & \\ e_1 & e_2 & & \\ f_1 & & & \\ g_1 & & & \end{array}$$

then

$$a_7 = c_3 = e_2 = g_1$$

and if the array is given by

$$\begin{array}{cccc} a_0 & a_2 & a_4 & a_6 \\ a_1 & a_3 & a_5 & 0 \\ b_1 & b_2 & b_3 & \\ c_1 & c_2 & 0 & \\ d_1 & d_2 & & \\ e_1 & 0 & & \\ f_1 & & & \end{array}$$

then

$$a_6 = b_3 = d_2 = f_1$$

In any case the last term of the first column is equal to a_n, or

$$a_n = \frac{\Delta_{n-1} a_n}{\Delta_{n-1}} = \frac{\Delta_n}{\Delta_{n-1}}$$

Thus it has been shown that the first column of the Routh array is given by

$$1, \quad \Delta_1, \quad \frac{\Delta_2}{\Delta_1}, \quad \frac{\Delta_3}{\Delta_2}, \quad \ldots, \quad \frac{\Delta_n}{\Delta_{n-1}}$$

Problem A-8-5 Show that if X and Y are functions of u and v; u and v are functions of x and y, then

$$\frac{\partial(X, Y)}{\partial(u, v)} \cdot \frac{\partial(u, v)}{\partial(x, y)} = \frac{\partial(X, Y)}{\partial(x, y)}$$

where

$$\frac{\partial(X, Y)}{\partial(x, y)} = \begin{vmatrix} \dfrac{\partial X}{\partial x} & \dfrac{\partial X}{\partial y} \\ \dfrac{\partial Y}{\partial x} & \dfrac{\partial Y}{\partial y} \end{vmatrix}$$

Solution

$$\frac{\partial(X, Y)}{\partial(u, v)} \cdot \frac{\partial(u, v)}{\partial(x, y)} = \begin{vmatrix} \dfrac{\partial X}{\partial u} & \dfrac{\partial X}{\partial v} \\ \dfrac{\partial Y}{\partial u} & \dfrac{\partial Y}{\partial v} \end{vmatrix} \cdot \begin{vmatrix} \dfrac{\partial u}{\partial x} & \dfrac{\partial u}{\partial y} \\ \dfrac{\partial v}{\partial x} & \dfrac{\partial v}{\partial y} \end{vmatrix}$$

$$= \begin{vmatrix} \dfrac{\partial X}{\partial u}\dfrac{\partial u}{\partial x} + \dfrac{\partial X}{\partial v}\dfrac{\partial v}{\partial x} & \dfrac{\partial X}{\partial u}\dfrac{\partial u}{\partial y} + \dfrac{\partial X}{\partial v}\dfrac{\partial v}{\partial y} \\ \dfrac{\partial Y}{\partial u}\dfrac{\partial u}{\partial x} + \dfrac{\partial Y}{\partial v}\dfrac{\partial v}{\partial x} & \dfrac{\partial Y}{\partial u}\dfrac{\partial u}{\partial y} + \dfrac{\partial Y}{\partial v}\dfrac{\partial v}{\partial y} \end{vmatrix}$$

$$= \begin{vmatrix} \dfrac{\partial X}{\partial x} & \dfrac{\partial X}{\partial y} \\ \dfrac{\partial Y}{\partial x} & \dfrac{\partial Y}{\partial y} \end{vmatrix} = \frac{\partial(X, Y)}{\partial(x, y)}$$

Problem A-8-6 Suppose that X and Y are functions of x and y. Show that if at least one of $\partial X/\partial x, \partial X/\partial y, \partial Y/\partial x, \partial Y/\partial y$ is nonzero and

$$\frac{\partial(X, Y)}{\partial(x, y)} = 0$$

then there exists a functional relationship between X and Y, namely, there exists a function f such that either

$$Y(x, y) = f(X(x, y)) \qquad \text{or} \qquad X(x, y) = f(Y(x, y))$$

Solution Without loss of generality we can assume that $\partial X/\partial x \neq 0$. $X = X(x, y)$ can be solved for x, or

$$x = h(X, y)$$

Then we obtain

$$Y = Y(h(X, y), y)$$

Thus X and Y are functions of X and y. Then

$$\frac{\partial(X, Y)}{\partial(x, y)} = \frac{\partial(X, Y)}{\partial(X, y)} \cdot \frac{\partial(X, y)}{\partial(x, y)} = \begin{vmatrix} 1 & 0 \\ \dfrac{\partial Y}{\partial X} & \dfrac{\partial Y}{\partial y} \end{vmatrix} \cdot \begin{vmatrix} \dfrac{\partial X}{\partial x} & \dfrac{\partial X}{\partial y} \\ 0 & 1 \end{vmatrix}$$

$$= \frac{\partial Y}{\partial y} \cdot \frac{\partial X}{\partial x}$$

Note that by assumption

$$\frac{\partial(X, Y)}{\partial(x, y)} = 0 \qquad \text{and} \qquad \frac{\partial X}{\partial x} \neq 0$$

Hence we obtain

$$\frac{\partial Y}{\partial y} = 0$$

From this we conclude that $Y = Y(h(X, y), y)$ does not involve y. Hence

$$Y = f(X)$$

Problem A-8-7 A general solution $F(x, y) = c$, where c is a constant, of the following first-order differential equation

$$y' = f(x, y)$$

is called an integral of this differential equation. Prove that there exists a functional relationship between two integrals F and G of the same first-order differential equation.

Solution Put

$$F(x, y) = a, \qquad G(x, y) = b$$

where a and b are arbitrary constants. Assume that y is a general solution of the differential equation corresponding to constants a and b. Differentiating F and G with respect to x, we obtain

$$\frac{\partial F}{\partial x} + \frac{\partial F}{\partial y} y' = 0$$

$$\frac{\partial G}{\partial x} + \frac{\partial G}{\partial y} y' = 0$$

Hence

$$\begin{vmatrix} \dfrac{\partial F}{\partial x} & \dfrac{\partial F}{\partial y} \\ \dfrac{\partial G}{\partial x} & \dfrac{\partial G}{\partial y} \end{vmatrix} = 0$$

This equation holds true for an arbitrary general solution $y(x)$ at point $(x, y(x))$. [If the initial condition is chosen properly, point $(x, y(x))$ can be an arbitrary point in some domain.] Hence from Prob. A-8-6, there exists a functional relationship between F and G.

Problem A-8-8 Suppose that X and Y are functions of x and y, namely $X = X(x, y)$ and $Y = Y(x, y)$. Prove that if at point (x_0, y_0)

$$\frac{\partial(X, Y)}{\partial(x, y)} \neq 0$$

then there exist functions $F(X, Y)$ and $G(X, Y)$ defined in the neighborhood of point $(X(x_0, y_0), Y(x_0, y_0))$ such that

$$x = F(X, Y), \qquad y = G(X, Y)$$

Solution By assumption we can assume $\partial X/\partial x \neq 0$ at point (x_0, y_0). Solving $X = X(x, y)$ for x in the neighborhood of point (x_0, y_0) we obtain

$$x = h(X, y)$$

Then, $Y = Y(h(X, y), y)$. Y is thus a function of X and y. Therefore, X and Y are functions of X and y; and X and y are functions of x and y. Hence, we obtain at point (x_0, y_0)

$$0 \neq \frac{\partial(X, Y)}{\partial(x, y)} = \frac{\partial(X, Y)}{\partial(X, y)} \cdot \frac{\partial(X, y)}{\partial(x, y)} = \frac{\partial Y}{\partial y} \cdot \frac{\partial X}{\partial x}$$

from which we obtain

$$\frac{\partial Y}{\partial y} \neq 0$$

We can, therefore, solve $Y = Y(h(X, y), y)$ for y and obtain $y = G(X, Y)$. Substituting $y = G(X, Y)$ into $x = h(X, y)$, we obtain

$$x = F(X, Y)$$

This completes the proof.

Problem A-8-9 (8-2) Assume that $\mathbf{x}(t)$ is the solution of the following equation:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, t), \qquad -\infty < t < \infty$$

corresponding to a fixed choice of **u**. Show that for fixed **u**, deviations from the motion $\mathbf{x}(t)$ are a free dynamic system.

Solution Let

$$\mathbf{y} = \mathbf{x} + \mathbf{z}$$

then

$$\begin{aligned}\dot{\mathbf{y}} &= \dot{\mathbf{x}} + \dot{\mathbf{z}} \\ &= \mathbf{f}(\mathbf{x} + \mathbf{z}, \mathbf{u}, t) \\ &= \mathbf{f}(\mathbf{x}, \mathbf{u}, t) + \mathbf{g}(\mathbf{z}, t)\end{aligned}$$

Since $\mathbf{x}(t)$ is a solution of the given equation it follows that

$$\dot{\mathbf{z}} = \mathbf{g}(\mathbf{z}, t)$$

$$\mathbf{g}(\mathbf{0}, t) = \mathbf{0} \qquad \text{for all } t$$

Hence, for fixed **u**, deviations from the motion $\mathbf{x}(t)$ are a free dynamic system.

Problem A-8-10 (*8-1*) Consider the following system:

$$\dot{x}_1 = a\epsilon x_1 - x_2 - ax_1^3$$

$$\dot{x}_2 = x_1 + a\epsilon x_2 - ax_2^3$$

where $a > 0$ and ϵ is a small positive constant. Show that the origin is unstable. The origin may, however, be regarded as "asymptotically stable" in the large, if we disregard small oscillations of amplitude less than $\sqrt{2\epsilon}$ around the origin.

Solution Near the origin the system can be linearized as follows:

$$\dot{x}_1 = a\epsilon x_1 - x_2$$

$$\dot{x}_2 = x_1 + a\epsilon x_2$$

The characteristic roots of this linearized system are

$$\lambda_1 = a\epsilon + j, \qquad \lambda_2 = a\epsilon - j$$

Since $a\epsilon > 0$, the origin is unstable.

We can, however, show that the origin may be regarded as if "asymptotically stable" in the large if we disregard small oscillations (a limit cycle) within a small neighborhood of the origin. Let us take

$$V = x_1^2 + x_2^2$$

then $\dot{V}$ can be obtained as

$$\dot{V} = -2a(x_1^4 + x_2^4) + 2a\epsilon(x_1^2 + x_2^2)$$

Outside the circle U with its center at the origin and radius $\sqrt{2\epsilon}$ we have

$$\sqrt{2\epsilon} < \sqrt{x_1^2 + x_2^2}$$

and

$$\dot{V} < -2a\left[(x_1^4 + x_2^4) - \frac{1}{2}(x_1^2 + x_2^2)^2\right] = -a(x_1^2 - x_2^2)^2 \leqslant 0$$

Hence outside the circle U, we have $V > 0$ and $\dot{V} < 0$. All solutions starting from the outside of the circle U approach a limit cycle located inside the circle U

as $t \to \infty$. Thus, the origin may be regarded as "asymptotically stable" in the large, if we disregard small oscillations of amplitude less than $\sqrt{2\epsilon}$ around the origin.

Problem A-8-11 Referring to Theorem 8–7 prove that the relationship

$$\mathbf{A}^*\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{Q}$$

is preserved under a transformation of coordinates.

Solution Let

$$\mathbf{x} = \mathbf{T}\mathbf{y}$$

where **T** is an $n \times n$ nonsingular matrix, then

$$\dot{\mathbf{y}} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}\mathbf{y} = \mathbf{A}_1\mathbf{y}$$

where

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \mathbf{A}_1$$

We thus have

$$\mathbf{A} = \mathbf{T}\mathbf{A}_1\mathbf{T}^{-1}$$

and

$$\mathbf{A}^*\mathbf{P} + \mathbf{P}\mathbf{A} = (\mathbf{T}\mathbf{A}_1\mathbf{T}^{-1})^*\mathbf{P} + \mathbf{P}(\mathbf{T}\mathbf{A}_1\mathbf{T}^{-1}) = -\mathbf{Q}$$

or

$$\mathbf{T}^*(\mathbf{T}^{-1})^*\mathbf{A}_1^*\mathbf{T}^*\mathbf{P}\mathbf{T} + \mathbf{T}^*\mathbf{P}\mathbf{T}\mathbf{A}_1\mathbf{T}^{-1}\mathbf{T} = -\mathbf{T}^*\mathbf{Q}\mathbf{T}$$

Therefore,

$$\mathbf{A}_1^*\mathbf{P}_1 + \mathbf{P}_1\mathbf{A}_1 = -\mathbf{Q}_1$$

where

$$\mathbf{P}_1 = \mathbf{T}^*\mathbf{P}\mathbf{T} \qquad \text{and} \qquad \mathbf{Q}_1 = \mathbf{T}^*\mathbf{Q}\mathbf{T}$$

Problem A-8-12 Determine the stability of the equilibrium state of the following system:

$$\begin{bmatrix}\dot{x}_1 \\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}-2 & -1-j \\ -1+j & -3\end{bmatrix}\begin{bmatrix}x_1 \\ x_2\end{bmatrix}$$

Solution The equilibrium state is the origin, or $\mathbf{x} = \mathbf{0}$. Solving the following equation:

$$\begin{bmatrix}-2 & -1-j \\ -1+j & -3\end{bmatrix}\begin{bmatrix}p_{11} & p_{12} \\ \bar{p}_{12} & p_{22}\end{bmatrix} + \begin{bmatrix}p_{11} & p_{12} \\ \bar{p}_{12} & p_{22}\end{bmatrix}\begin{bmatrix}-2 & -1-j \\ -1+j & -3\end{bmatrix} = \begin{bmatrix}-1 & 0 \\ 0 & -1\end{bmatrix}$$

we obtain:

$$\begin{aligned}4p_{11} + (1-j)p_{12} + (1+j)\bar{p}_{12} &= 1 \\ (1-j)p_{11} + 5\bar{p}_{12} + (1-j)p_{22} &= 0 \\ (1+j)p_{11} + 5p_{12} + (1+j)p_{22} &= 0 \\ (1-j)p_{12} + (1+j)\bar{p}_{12} + 6p_{22} &= 1\end{aligned} \tag{8-53}$$

Notice that since p_{11} and p_{22} are real, the second and the third of Eq. (8–53) are equivalent. They are conjugate to each other. From Eq. (8–53) we obtain

$$p_{11} = \frac{3}{8}, \qquad p_{12} = -\frac{1}{8}(1 + j), \qquad p_{22} = \frac{1}{4}$$

or

$$\mathbf{P} = \begin{bmatrix} \frac{3}{8} & -\frac{1}{8} - j\frac{1}{8} \\ -\frac{1}{8} + j\frac{1}{8} & \frac{1}{4} \end{bmatrix}$$

which is positive definite. Hence we conclude that the origin of the system is asymptotically stable.

Problem A-8-13 Show that if

$$\hat{\mathbf{F}}(\mathbf{x}) = \mathbf{F}^*(\mathbf{x}) + \mathbf{F}(\mathbf{x})$$

where

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

is negative definite, then the determinant of $\mathbf{F}(\mathbf{x})$ is nonzero for $\mathbf{x} \neq \mathbf{0}$, or

$$\det \mathbf{F}(\mathbf{x}) \neq 0 \qquad \text{for } \mathbf{x} \neq \mathbf{0}$$

Solution By use of the relationship

$$\mathbf{y}^*\mathbf{F}(\mathbf{x})\mathbf{y} = [\mathbf{y}^*\mathbf{F}(\mathbf{x})\mathbf{y}]^*$$

we have

$$\begin{aligned} \mathbf{y}^*\hat{\mathbf{F}}(\mathbf{x})\mathbf{y} &= \mathbf{y}^*[\mathbf{F}^*(\mathbf{x}) + \mathbf{F}(\mathbf{x})]\mathbf{y} \\ &= \mathbf{y}^*\mathbf{F}^*(\mathbf{x})\mathbf{y} + \mathbf{y}^*\mathbf{F}(\mathbf{x})\mathbf{y} \\ &= [\mathbf{y}^*\mathbf{F}(\mathbf{x})\mathbf{y}]^* + \mathbf{y}^*\mathbf{F}(\mathbf{x})\mathbf{y} \\ &= 2\mathbf{y}^*\mathbf{F}(\mathbf{x})\mathbf{y} \end{aligned}$$

Since $\hat{\mathbf{F}}(\mathbf{x})$ is negative definite, $\mathbf{F}(\mathbf{x})$ is also negative definite. Hence

$$\det \mathbf{F}(\mathbf{x}) \neq 0 \qquad \text{for } \mathbf{x} \neq \mathbf{0}$$

Problem A-8-14 By the use of Krasovskii's theorem (Theorem 8–13) prove that the equilibrium state $\mathbf{x} = \mathbf{0}$ of the system defined by

$$\dot{x}_1 = -3x_1 + x_2$$
$$\dot{x}_2 = x_1 - x_2 - x_2^3$$

is asymptotically stable in the large.

Solution In this problem

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} -3x_1 + x_2 \\ x_1 - x_2 - x_2^3 \end{bmatrix}$$

and

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} -3 & 1 \\ 1 & -1 - 3x_2^2 \end{bmatrix}$$

Hence

$$\begin{aligned} \hat{\mathbf{F}}(\mathbf{x}) &= \mathbf{F}'(\mathbf{x}) + \mathbf{F}(\mathbf{x}) \\ &= \begin{bmatrix} -6 & 2 \\ 2 & -2 - 6x_2^2 \end{bmatrix} \end{aligned}$$

Since $\hat{\mathbf{F}}(\mathbf{x})$ is negative definite, the equilibrium state $\mathbf{x} = \mathbf{0}$ is asymptotically stable. In addition,

$$\mathbf{f}'(\mathbf{x})\mathbf{f}(\mathbf{x}) = (-3x_1 + x_2)^2 + (x_1 - x_2 - x_2^3)^2 \to \infty \qquad \text{as } \|\mathbf{x}\| \to \infty$$

The equilibrium state is, therefore, asymptotically stable in the large.

Problem A-8-15 (8–20) Prove that the function $V(x, y)$, the solution to the following partial differential equation,

$$\frac{\partial V}{\partial x} f_1(x, y) + \frac{\partial V}{\partial y} f_2(x, y) = -\phi(x, y)[1 - V(x, y)] \tag{8–54}$$

where $\phi(x, y)$ is a quadratic form and is positive definite, is a Liapunov function establishing the asymptotic stability of the origin $x = y = 0$ of the following system:

$$\begin{aligned} \dot{x} &= f_1(x, y) \\ \dot{y} &= f_2(x, y) \end{aligned} \tag{8–55}$$

where the equations of the linear approximation are asymptotically stable. It is assumed that the solution to Eq. (8–55) is continuous for $-\infty < t < \infty$. (This is a simplified version of Zubov's method of construction of Liapunov functions.)

Solution Because the roots of the characteristic equation of the linearized system have negative real parts, the leading term of the series expansion of $V(x, y)$ is a positive definite quadratic form. Thus, $V(x, y)$ is positive definite in some neighborhood of the origin and also $V(0, 0) = 0$. Equation (8–54) gives the derivative of $V(x, y)$ along the integral curves of the system, Eq. (8–55), and since, near the origin, Eq. (8–54) reduces to

$$\frac{dV}{dt} = -\phi(x, y)$$

we see that dV/dt is negative definite in the neighborhood of the origin. In fact, dV/dt is negative whenever $0 \leqslant V < 1$. From Theorem 8–2 $V(x, y)$ is a Liapunov function establishing the asymptotic stability of the origin.

Problem A-8-16 (8-2) E. A. Barbashin obtained a specific stability criterion for the third-order system given by the following equations:

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 \\ \dot{x}_3 &= -f(x_1) - g(x_2) - a_1 x_3 \end{aligned} \tag{8-56}$$

where $f(0) = g(0) = 0$ and $f(x_1)$ and $g(x_2)$ are differentiable.

Prove that the equilibrium state $\mathbf{x} = \mathbf{0}$ of the system is asymptotically stable in the large if

(1) $a_1 > 0$

(2) $f'(x_1) \geqslant \epsilon_1 > 0 \qquad$ for $x_1 \neq 0$

(3) $\dfrac{a_1 g(x_2)}{x_2} - f'(x_1) > 0 \qquad$ for $x_2 \neq 0$

where

$$f'(x_1) = \frac{d}{dx_1} f(x_1)$$

and a Liapunov function may be written as

$$V(\mathbf{x}) = a_1 F(x_1) + f(x_1)x_2 + G(x_2) + \tfrac{1}{2}(a_1 x_2 + x_3)^2 \tag{8-57}$$

where

$$F(x_1) = \int_0^{x_1} f(x_1)\, dx_1 \qquad \text{and} \qquad G(x_2) = \int_0^{x_2} g(x_2)\, dx_2$$

Solution In view of (2), $f'(0) > 0$ and $F(x_1) > 0$ for $x_1 \neq 0$. Further, by (1) and (3) it follows that $G(x_2) > 0$ for $x_2 \neq 0$. For $x_2 = 0$, Eq. (8-57) is nonnegative and is 0 only if $x_1 = x_3 = 0$. For $x_2 \neq 0$, Eq. (8-57) becomes as follows:

$$\begin{aligned} V(\mathbf{x}) = {} & \frac{[2G(x_2) + f(x_1)x_2]^2}{4G(x_2)} \\ & + \frac{\int_0^{x_1} f(x_1) \int_0^{x_2} \left\{a_1[g(x_2)/x_2] - f'(x_1)\right\} x_2\, dx_2\, dx_1}{G(x_2)} \\ & + \frac{(a_1 x_2 + x_3)^2}{2} \end{aligned}$$

The first and third terms in the preceding expression for $V(\mathbf{x})$ are clearly nonnegative. The second term is positive because the inner integral is positive and $F(x_1) > 0$ for $x_1 \neq 0$. Hence,

$$V(\mathbf{x}) > 0 \qquad \text{for } \mathbf{x} \neq \mathbf{0}$$

and

$$V(\mathbf{0}) = 0$$

The time derivative of V is found as

$$\dot{V}(\mathbf{x}) = -\left[\frac{a_1 g(x_2)}{x_2} - f'(x_1)\right] x_2^2$$

Because of (3)

$$\begin{aligned} \dot{V}(\mathbf{x}) &\leqslant 0 \qquad \text{for all } \mathbf{x} \neq \mathbf{0} \\ \dot{V}(\mathbf{0}) &= 0 \end{aligned}$$

It is clearly seen that $\dot{V}(\mathbf{x})$ does not vanish identically along any trajectory starting at $\mathbf{x}_0 \neq \mathbf{0}$. [Since $\dot{V}(\mathbf{x}) = 0$ only if $x_2 = 0$, we see from Eq. (8-56) that $\dot{V}(\mathbf{x})$ identically equal to zero corresponds to $x_1 = x_2 = x_3 = 0$.] Hence, Eq. (8-57) is a Liapunov function for the system of Eq. (8-56). In addition, we see that

$$V(\mathbf{x}) \to \infty \qquad \text{as } \|\mathbf{x}\| \to \infty$$

Hence by Theorem 8-5, the equilibrium state $\mathbf{x} = \mathbf{0}$ of the system is asymptotically stable in the large.

Notice that Eq. (8-56) is equivalent to

$$\dddot{x}_1 + a_1 \ddot{x}_1 + g(\dot{x}_1) + f(x_1) = 0$$

If we let $f(x_1) = a_3 x_1$, and $g(\dot{x}_1) = a_2 \dot{x}_1$, then this last equation becomes as follows:

$$\dddot{x}_1 + a_1 \ddot{x}_1 + a_2 \dot{x}_1 + a_3 x_1 = 0$$

The three conditions for asymptotic stability are

(1) $a_1 > 0$

(2) $a_3 > 0$

(3) $a_1 a_2 - a_3 > 0$

(Note that these three conditions imply $a_2 > 0$.) Notice that these conditions are simply the Routh-Hurwitz conditions for asymptotic stability of the linearized third order system.

Problem A-8-17 (8-20) By use of the variable gradient method construct a Liapunov function for the following system:

$$\begin{aligned} \dot{x}_1 &= -x_1 + 2x_1^2 x_2 \\ \dot{x}_2 &= -x_2 \end{aligned}$$

Solution Let the gradient of V be defined by

$$\nabla V = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + 2x_2 \end{bmatrix}$$

The derivative of V is then given by

$$\begin{aligned} \dot{V} &= (\nabla V)' \dot{\mathbf{x}} \\ &= (a_{11}x_1 + a_{12}x_2)\dot{x}_1 + (a_{21}x_1 + 2x_2)\dot{x}_2 \\ &= -a_{11}x_1^2 + 2a_{11}x_1^3 x_2 - a_{12}x_1 x_2 \\ &\quad + 2a_{12}x_1^2 x_2^2 - a_{21}x_1 x_2 - 2x_2^2 \end{aligned} \tag{8-58}$$

Let us put, as a trial

$$a_{11} = 1 \qquad \text{and} \qquad a_{12} = a_{21} = 0$$

Then Eq. (8–58) becomes

$$\dot{V} = -x_1^2(1 - 2x_1x_2) - 2x_2^2$$

$\dot{V}$ is negative definite if

$$1 - 2x_1x_2 > 0 \tag{8-59}$$

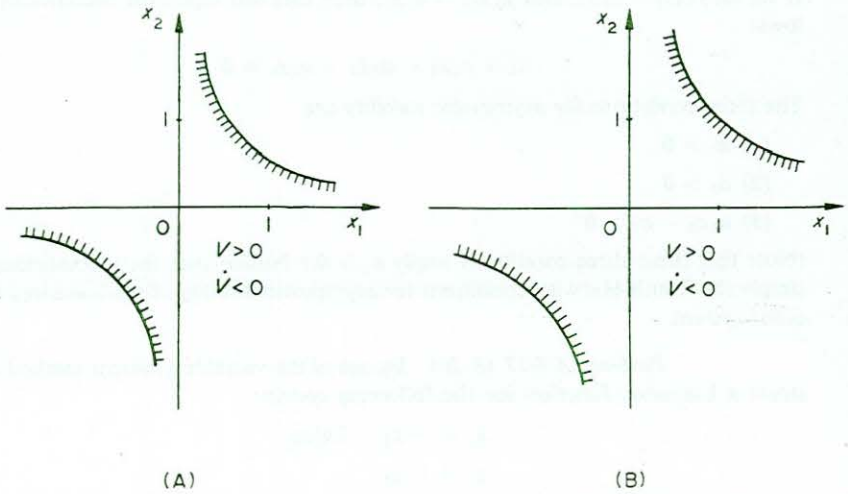
Thus, Eq. (8–59) is a constraint on x_1 and x_2. The gradient ∇V becomes

$$\nabla V = \begin{bmatrix} x_1 + 0 \\ 0 + 2x_2 \end{bmatrix}$$

We see that the curl equation is satisfied. V is obtained as

$$V = \int_0^{x_1,\ (x_2=0)} x_1 dx_1 + \int_0^{x_2,\ (x_1=x_1)} 2x_2 dx_2$$
$$= \frac{x_1^2}{2} + x_2^2$$

V is clearly positive definite. From this V function, we can say that the origin of the system is asymptotically stable. Figure 8–8(A) is a plot showing the region for which V is positive definite and $\dot{V}$ is negative definite.



(A) (B)

Figure 8-8 Regions for which V is positive definite and $\dot{V}$ is negative definite.
(A) $V = \frac{x_1^2}{2} + x_2^2$; (B) $V = \frac{x_1^2}{1 - x_1x_2} + x_2^2$.

In order to illustrate how the choice of a_{11}, a_{12}, and a_{21} affects the resulting V function, let us choose

$$a_{11} = \frac{2}{(1 - x_1x_2)^2}, \qquad a_{12} = \frac{-x_1^2}{(1 - x_1x_2)^2}, \qquad a_{21} = \frac{x_1^2}{(1 - x_1x_2)^2}$$

Then,

$$\dot{V} = -2x_1^2 - 2x_2^2$$

$\dot{V}$ is negative definite in the entire state plane. ∇V becomes

$$\nabla V = \begin{bmatrix} \dfrac{2x_1}{(1 - x_1x_2)^2} - \dfrac{x_1^2x_2}{(1 - x_1x_2)^2} \\ \dfrac{x_1^3}{(1 - x_1x_2)^2} + 2x_2 \end{bmatrix}$$

It follows that

$$\frac{\partial \nabla V_2}{\partial x_1} = \frac{3x_1^2 - x_1^3x_2}{(1 - x_1x_2)^3}$$
$$\frac{\partial \nabla V_1}{\partial x_2} = \frac{3x_1^2 - x_1^3x_2}{(1 - x_1x_2)^3}$$

Thus the condition for the curl equation is satisfied: V is obtained as

$$V = \int_0^{x_1,\ (x_2=0)} \left[\frac{2x_1}{(1 - x_1x_2)^2} - \frac{x_1^2x_2}{(1 - x_1x_2)^2}\right] dx_1 + \int_0^{x_2,\ (x_1=x_1)} \left[\frac{x_1^3}{(1 - x_1x_2)^2} + 2x_2\right] dx_2$$
$$= \frac{x_1^2}{1 - x_1x_2} + x_2^2 \tag{8-60}$$

From Eq. (8–60) V is positive definite if $1 - x_1x_2 > 0$. The region

$$1 > x_1x_2 \tag{8-61}$$

includes the origin. From this V function we can say that the origin of the system is asymptotically stable. Figure 8–8(B) shows a plot of the region given by inequality (8–61). This is the region for which V is positive definite and $\dot{V}$ is negative definite.

It is seen that the second V function gives a larger region, for which V is positive definite and $\dot{V}$ is negative definite, than the first V function does. Hence

$$V = \frac{x_1^2}{1 - x_1x_2} + x_2^2$$

is a better choice between the two V functions.

Problem A-8-18 (8–21) Prove that the origin of the following system is unstable:

$$\dot{x}_1 = -2x_1x_2x_3$$

$$\dot{x}_2 = \frac{x_2}{t + 2} + x_1^2x_3$$

$$\dot{x}_3 = \frac{x_3}{t + 2} + x_1^2x_2$$

where $t \geqslant 0$.

Solution The equilibrium states of the given system can be found by equating $\dot{x}_1$, $\dot{x}_2$, and $\dot{x}_3$ to zero. Clearly, they are given by $(x_1, 0, 0)$, x_1 being any constant. We shall examine stability of the origin. After some inspection, choose

$$\nabla V = \begin{bmatrix} 2x_1 \\ 2x_2 + 2\left(\frac{t+1}{t+2}\right)x_3 \\ 2\left(\frac{t+1}{t+2}\right)x_2 + 2x_3 \end{bmatrix}$$

then

$$\int_0^{\mathbf{x}} (\nabla V)'d\mathbf{x} = \int_0^{x_1,\ (x_2=x_3=0)} 2x_1 dx_1 + \int_0^{x_2,\ (x_1=x_1,\ x_3=0)} 2x_2 dx_2$$
$$+ \int_0^{x_3,\ (x_1=x_1,\ x_2=x_2)} \left(2\frac{t+1}{t+2}x_2 + 2x_3\right) dx_3$$
$$= x_1^2 + x_2^2 + x_3^2 + \frac{2(t+1)}{t+2}x_2x_3$$

If we choose

$$V = x_1^2 + x_2^2 + x_3^2 + \frac{2(t+1)}{t+2}x_2x_3$$
$$= x_1^2 + \frac{1}{t+2}(x_2^2 + x_3^2) + \frac{t+1}{t+2}(x_2 + x_3)^2$$

then V is positive definite for $t \geqslant 0$. Since this V function is time-varying, $\dot{V}$ is not equal to $(\nabla V)'\dot{\mathbf{x}}$. $\dot{V}$ is given by

$$\dot{V} = \frac{\partial V}{\partial t} + (\nabla V)'\dot{\mathbf{x}}$$
$$= \frac{2}{t+2}\left[x_2^2 + x_3^2 + \frac{2t+3}{t+2}x_2x_3 + x_1^2(t+1)(x_2^2 + x_3^2)\right]$$
$$= \frac{2}{t+2}\left\{\left[\frac{0.5}{t+2} + (t+1)x_1^2\right](x_2^2 + x_3^2) + \frac{t+1.5}{t+2}(x_2 + x_3)^2\right\}$$

Thus $\dot{V} > 0$ except on the x_1 axis where $\dot{V} = 0$. By Theorem 8–6 the origin of the system is unstable.

Problem A-8-19 (8–22) Find conditions for stability of the equilibrium state of the following system:

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = a_1(t)x_1 + a_2(t)x_2$$

Solution The equilibrium state is clearly the origin. Let us assume

$$\nabla V = \begin{bmatrix} b_{11}x_1 + b_{12}x_2 \\ b_{21}x_1 + b_{22}x_2 \end{bmatrix}$$

then

$$(\nabla V)'\dot{\mathbf{x}} = b_{11}x_1x_2 + b_{12}x_2^2 + b_{21}a_1(t)x_1^2 + b_{22}a_2(t)x_2^2$$
$$+ b_{21}a_2(t)x_1x_2 + b_{22}a_1(t)x_1x_2$$

Let us choose $b_{12} = b_{21} = 0$, then

$$(\nabla V)'\dot{\mathbf{x}} = b_{11}x_1x_2 + b_{22}a_2(t)x_2^2 + b_{22}a_1(t)x_1x_2$$

Let

$$b_{11} + b_{22}a_1(t) = 0$$

By choosing $b_{11} = 2$, and assuming $a_1(t) \neq 0$, we obtain

$$b_{22} = -\frac{2}{a_1(t)}$$

∇V can then be written as

$$\nabla V = \begin{bmatrix} 2x_1 \\ -\frac{2}{a_1(t)}x_2 \end{bmatrix}$$

Since

$$\int_0^{x_1,\ (x_2=0)} 2x_1 dx_1 + \int_0^{x_2,\ (x_1=x_1)} \frac{-2}{a_1(t)}x_2 dx_2 = x_1^2 - \frac{1}{a_1(t)}x_2^2$$

let us choose

$$V = x_1^2 - \frac{1}{a_1(t)}x_2^2$$

Then

$$\dot{V} = 2x_1\dot{x}_1 + \frac{\dot{a}_1(t)}{[a_1(t)]^2}x_2^2 - \frac{1}{a_1(t)}2x_2\dot{x}_2$$
$$= -\left(\frac{2a_2(t)}{a_1(t)} - \frac{\dot{a}_1(t)}{a_1^2(t)}\right)x_2^2$$

If $a_1(t) < 0$ for $t \geqslant t_0$, then V is positive definite. If

$$-\frac{2a_2(t)}{a_1(t)} + \frac{\dot{a}_1(t)}{a_1^2(t)} < 0 \qquad \text{for } t \geqslant t_0$$

then $\dot{V}$ is negative semidefinite. Hence, by Theorem 8–3 conditions for stability may be obtained as follows:

$$a_1(t) < 0 \qquad \text{for } t \geqslant t_0$$
$$-\frac{2a_2(t)}{a_1(t)} + \frac{\dot{a}_1(t)}{a_1^2(t)} < 0 \qquad \text{for } t \geqslant t_0$$

Problem A-8-20 (8–23, 8–24) Consider the problem of stability of a homogeneous atomic reactor with constant power extraction. The reactor equations are

$$\frac{d}{dt}\log P = -\frac{\alpha}{\tau}T$$
$$\frac{d}{dt}T = (P-1)\frac{1}{\epsilon} \qquad (8\text{–}62)$$

where $P(t)$ is the instantaneous reactor power, $T(t)$ is the temperature, α the temperature coefficient, τ the average life of a neutron, and ϵ the heat capacity. Equations (8–62) can be rewritten as

$$\dot{z}_1 = -\frac{\alpha}{\tau} z_2$$

$$\dot{z}_2 = \frac{1}{\epsilon}(e^{z_1} - 1) = f(z_1)$$

These are obtained by applying the transformation of coordinates

$$\log P = z_1, \qquad T = z_2$$

to the system, Eq. (8–62). Clearly the origin $z_1 = z_2 = 0$ is the equilibrium state. Examine stability of the origin by constructing a Liapunov function by means of the variable gradient method.

Solution Let the gradient of V be defined by

$$\nabla V = \begin{bmatrix} a_{11} z_1 + a_{12} z_2 \\ a_{21} z_1 + z_2 \end{bmatrix}$$

The values of the a's must be such that the following curl equation is satisfied:

$$\frac{\partial \nabla V_1}{\partial z_2} = \frac{\partial \nabla V_2}{\partial z_1}$$

or

$$z_1 \frac{\partial a_{11}}{\partial z_2} + \frac{\partial (a_{12} z_2)}{\partial z_2} = \frac{\partial (a_{21} z_1)}{\partial z_1}$$

We may choose $a_{12} = a_{21} = 0$ and $a_{11} = a_{11}(z_1)$. Then

$$\dot{V} = (\nabla V)' \dot{\mathbf{z}}$$

$$= [a_{11}(z_1)] z_1 \left(-\frac{\alpha}{\tau} z_2\right) + z_2 \frac{1}{\epsilon}(e^{z_1} - 1)$$

If we choose

$$a_{11}(z_1) = \frac{1}{z_1}\left(\frac{\tau}{\alpha\epsilon}\right)(e^{z_1} - 1)$$

then,

$$\dot{V} = -\frac{z_2}{\epsilon}(e^{z_1} - 1) + \frac{z_2}{\epsilon}(e^{z_1} - 1) = 0$$

With this choice of $a_{11}(z_1)$, ∇V can be written as

$$\nabla V = \begin{bmatrix} \frac{\tau}{\alpha\epsilon}(e^{z_1} - 1) \\ z_2 \end{bmatrix}$$

V is found by line integration as follows:

$$V = \int_0^{z_1,\ (z_2=0)} \left(\frac{\tau}{\alpha\epsilon}\right)(e^{z_1} - 1) dz_1 + \int_0^{z_2,\ (z_1=z_1)} z_2 dz_2$$

$$= \frac{\tau}{\alpha\epsilon}(e^{z_1} - z_1 - 1) + \frac{z_2^2}{2}$$

Since τ and ϵ are positive quantities, if α is positive, then

$$V > 0 \qquad \text{for } z_1, z_2 \neq 0$$

$$V = 0 \qquad \text{for } z_1 = z_2 = 0$$

Thus V is a positive definite scalar function. We conclude that the equilibrium state, the origin, is stable. From this it follows that the solution $P = 1$, $T = 0$ of the system of Eq. (8–62) is stable for $\tau > 0$, $\epsilon > 0$, and $\alpha > 0$.

Problem A-8-21 (8–10) Determine the stability of the origin of the following system:

$$\dot{x}_1 = x_2$$

$$\dot{x}_2 = -x_2 - x_1^3$$

Solution An example of possible Liapunov functions is

$$V = x_1^4 + x_1^2 + 2x_1 x_2 + 2x_2^2$$

Its time derivative is

$$\dot{V} = -2x_1^4 - 2x_2^2$$

The origin of the system is clearly asymptotically stable in the large.

Problem A-8-22 (8–20) Determine the stability of the origin of the following system:

$$\dot{x}_1 = -2x_1 + 2x_2^4$$

$$\dot{x}_2 = -x_2$$

Solution An example of Liapunov functions is

$$V = 6x_1^2 + 12x_2^2 + 4x_1 x_2^4 + x_2^8$$

Its time derivative is

$$\dot{V} = -24(x_1^2 + x_2^2)$$

The origin of the system is clearly asymptotically stable in the large.

Problem A-8-23 (8–20) Determine the stability of the origin of the following system:

$$\dot{x}_1 = -x_1 + x_2 + x_1(x_1^2 + x_2^2)$$

$$\dot{x}_2 = -x_1 - x_2 + x_2(x_1^2 + x_2^2)$$

Solution An example of possible Liapunov functions is

$$V = x_1^2 + x_2^2$$

The time derivative of V is

$$\dot{V} = -2x_1^2 - 2x_2^2 + 2(x_1^2 + x_2^2)^2$$

Hence, the region for which $\dot{V}$ is negative definite is the inside of the circle defined by $x_1^2 + x_2^2 = 1$. From this V function we can conclude that the origin of the system is asymptotically stable.

Problem A-8-24 Determine stability of the origin of the following system:

$$\dot{x}_1 = \frac{-6x_1}{(1 + x_1^2)^2} + 2x_2$$

$$\dot{x}_2 = \frac{-2x_1}{(1 + x_1^2)^2} - \frac{2x_2}{(1 + x_1^2)^2}$$

Solution An example of Liapunov functions for this system is

$$V = \frac{x_1^2}{1 + x_1^2} + x_2^2$$

and $\dot{V}$ is given by

$$\dot{V} = -\frac{12x_1^2}{(1 + x_1^2)^4} - \frac{4x_2^2}{(1 + x_1^2)^2}$$

The origin of the system is asymptotically stable.

Problem A-8-25 (8-10) Obtain conditions for stability of the origin of the following system:

$$\dot{x}_1 = x_2$$

$$\dot{x}_2 = -b(x_1)x_1 - a(x_1, x_2)x_2$$

Solution A Liapunov function for this system is

$$V = 2\int_0^{x_1} b(x_1)x_1\,dx_1 + x_2^2$$

and its time derivative is

$$\dot{V} = -2a(x_1, x_2)x_2^2$$

Conditions for stability obtainable from this Liapunov function are

$$2\int_0^{x_1} b(x_1)x_1\,dx_1 + x_2^2 > 0, \qquad a(x_1, x_2) \geqslant 0$$

Problem A-8-26 (8-11) Obtain conditions for asymptotic stability of the origin of the following system:

$$\dot{x}_1 = x_2$$

$$\dot{x}_2 = x_3$$

$$\dot{x}_3 = -a_2x_2 - a_1x_3 - f(x_1)x_1$$

Solution An example of possible Liapunov functions is

$$V = \frac{a_2^2x_1^2}{2} + a_1\int_0^{x_1} f(x_1)x_1\,dx_1 + a_1a_2x_1x_2 + (a_1^2 + a_2)\frac{x_2^2}{2} + a_2x_1x_3 + a_1x_2x_3 + x_3^2$$

$$= \left(\frac{a_2x_1}{\sqrt{2}} + \frac{a_1x_2}{\sqrt{2}} + \frac{x_3}{\sqrt{2}}\right)^2 + a_1\int_0^{x_1} f(x_1)x_1\,dx_1 + a_2\frac{x_2^2}{2} + \frac{x_3^2}{2}$$

and its time derivative is

$$\dot{V} = -[a_2f(x_1)x_1^2 + 2f(x_1)x_1x_3 + a_1x_3^2]$$

Conditions for asymptotic stability obtainable from this Liapunov function are

$$a_1 > 0, \qquad a_2 > 0, \qquad f(x_1) > 0, \qquad a_1a_2 - f(x_1) > 0$$

Problem A-8-27 Show that the following V function is positive definite: [Refer to Eq. (8-39).]

$$V = \mathbf{x}^*\mathbf{S}\mathbf{x} + \int_0^e h(e)\,de$$

where

$$\int_0^e h(e)\,de \geqslant 0 \qquad [e \text{ is defined by Eq. (8-32).}]$$

$\mathbf{x} = n$ vector

$$\mathbf{S} = \begin{bmatrix} -\frac{\bar{a}_1a_1}{\bar{\lambda}_1 + \lambda_1} & -\frac{\bar{a}_1a_2}{\bar{\lambda}_1 + \lambda_2} & \cdots & -\frac{\bar{a}_1a_n}{\bar{\lambda}_1 + \lambda_n} \\ -\frac{\bar{a}_2a_1}{\bar{\lambda}_2 + \lambda_1} & -\frac{\bar{a}_2a_2}{\bar{\lambda}_2 + \lambda_2} & \cdots & -\frac{\bar{a}_2a_n}{\bar{\lambda}_2 + \lambda_n} \\ \vdots & \vdots & & \vdots \\ -\frac{\bar{a}_na_1}{\bar{\lambda}_n + \lambda_1} & -\frac{\bar{a}_na_2}{\bar{\lambda}_n + \lambda_2} & \cdots & -\frac{\bar{a}_na_n}{\bar{\lambda}_n + \lambda_n} \end{bmatrix}$$

and each λ_i possesses a negative real part.

Show also that

$$\mathbf{A}^*\mathbf{S} + \mathbf{S}\mathbf{A}$$

is negative semidefinite, where

$$\mathbf{A} = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix}$$

Solution Since the second term of V is positive semidefinite, we need to prove that $\mathbf{x}^*\mathbf{S}\mathbf{x}$ is positive definite. Notice that since each λ_i possesses a negative real part, we obtain

$$-\frac{\bar{a}_i a_k}{\bar{\lambda}_i + \lambda_k} = \int_0^\infty (\bar{a}_i e^{\bar{\lambda}_i t})(a_k e^{\lambda_k t})\, dt$$

Hence $\mathbf{x}^*\mathbf{S}\mathbf{x}$ can be written as

$$\begin{aligned}\mathbf{x}^*\mathbf{S}\mathbf{x} &= \sum_{i=1}^n \sum_{k=1}^n \int_0^\infty (\bar{a}_i e^{\bar{\lambda}_i t})(a_k e^{\lambda_k t})\bar{x}_i x_k\, dt \\ &= \int_0^\infty \sum_{i=1}^n (\bar{a}_i e^{\bar{\lambda}_i t}\bar{x}_i) \sum_{k=1}^n (a_k e^{\lambda_k t} x_k)\, dt \\ &= \int_0^\infty |\sum_{i=1}^n a_i e^{\lambda_i t} x_i|^2 dt > 0 \qquad \text{for } \mathbf{x} \neq \mathbf{0} \\ &= 0 \qquad \text{for } \mathbf{x} = \mathbf{0}\end{aligned}$$

Thus $\mathbf{x}^*\mathbf{S}\mathbf{x}$ is positive definite. It follows that V is also positive definite.

Next, we shall show that $\mathbf{A}^*\mathbf{S} + \mathbf{S}\mathbf{A}$ is negative semidefinite. Notice that

$$\mathbf{A}^*\mathbf{S} = -\begin{bmatrix} \frac{\bar{\lambda}_1}{\bar{\lambda}_1 + \lambda_1}\bar{a}_1 a_1 & \frac{\bar{\lambda}_1}{\bar{\lambda}_1 + \lambda_2}\bar{a}_1 a_2 & \cdots & \frac{\bar{\lambda}_1}{\bar{\lambda}_1 + \lambda_n}\bar{a}_1 a_n \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \frac{\bar{\lambda}_n}{\bar{\lambda}_n + \lambda_1}\bar{a}_n a_1 & \frac{\bar{\lambda}_n}{\bar{\lambda}_n + \lambda_2}\bar{a}_n a_2 & \cdots & \frac{\bar{\lambda}_n}{\bar{\lambda}_n + \lambda_n}\bar{a}_n a_n \end{bmatrix}$$

$$\mathbf{S}\mathbf{A} = -\begin{bmatrix} \frac{\lambda_1}{\bar{\lambda}_1 + \lambda_1}\bar{a}_1 a_1 & \frac{\lambda_2}{\bar{\lambda}_1 + \lambda_2}\bar{a}_1 a_2 & \cdots & \frac{\lambda_n}{\bar{\lambda}_1 + \lambda_n}\bar{a}_1 a_n \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \frac{\lambda_1}{\bar{\lambda}_n + \lambda_1}\bar{a}_n a_1 & \frac{\lambda_2}{\bar{\lambda}_n + \lambda_2}\bar{a}_n a_2 & \cdots & \frac{\lambda_n}{\bar{\lambda}_n + \lambda_n}\bar{a}_n a_n \end{bmatrix}$$

Therefore, we obtain

$$\mathbf{A}^*\mathbf{S} + \mathbf{S}\mathbf{A} = -\begin{bmatrix} \bar{a}_1 a_1 & \bar{a}_1 a_2 & \cdots & \bar{a}_1 a_n \\ \bar{a}_2 a_1 & \bar{a}_2 a_2 & \cdots & \bar{a}_2 a_n \\ \vdots & \vdots & & \vdots \\ \bar{a}_n a_1 & \bar{a}_n a_2 & \cdots & \bar{a}_n a_n \end{bmatrix}$$

Let us put

$$\mathbf{Q} = -(\mathbf{A}^*\mathbf{S} + \mathbf{S}\mathbf{A})$$

Then,

$$\mathbf{x}^*\mathbf{Q}\mathbf{x} = \sum_{i=1}^n \sum_{k=1}^n (\bar{a}_i a_k)\bar{x}_i x_k = \left|\sum_{i=1}^n (a_i x_i)\right|^2 \geqslant 0$$

Hence $\mathbf{Q}$ is positive semidefinite. Therefore, $\mathbf{A}^*\mathbf{S} + \mathbf{S}\mathbf{A}$ is negative semidefinite.

Problem A-8-28 Consider the system governed by the difference equation

$$\mathbf{x}((k+1)T) = \mathbf{f}(\mathbf{x}(kT), \mathbf{u}(kT))$$

Assuming that $\mathbf{u}$ is a constant vector, or $\mathbf{u} = \mathbf{c}$ (where $\mathbf{c}$ is a constant vector), find the equilibrium state $\mathbf{x}_e$.

Solution The equilibrium state has the property that if the initial state is $\mathbf{x}(0) = \mathbf{x}_e$, then repeated iterations of the transformation do not result in any change of state. Hence,

$$\mathbf{x}(nT) = \mathbf{x}_e$$

for all $n \geqslant 0$. Thus $\mathbf{x}_e$ can be found from

$$\mathbf{x}_e = \mathbf{f}(\mathbf{x}_e, \mathbf{c})$$

Obviously, the equilibrium state is a function of $\mathbf{c}$.

Problem A-8-29 (*8-2*) Consider the system

$$\mathbf{x}((k+1)T) = \mathbf{H}(\mathbf{x}(kT))\mathbf{x}(kT)$$

Assume that there are positive constants $c_1, c_2, \ldots, c_n$ such that either

(a) $$\max_i \left\{\sum_{j=1}^n \frac{c_i}{c_j}\left|h_{ij}(\mathbf{x})\right|\right\} < 1 \qquad \text{for all } \mathbf{x}$$

or

(b) $$\max_j \left\{\sum_{i=1}^n \frac{c_i}{c_j}\left|h_{ij}(\mathbf{x})\right|\right\} < 1 \qquad \text{for all } \mathbf{x}$$

Show that in either case $\mathbf{H}(\mathbf{x})\mathbf{x}$ is a contraction for all $\mathbf{x}$ and therefore the equilibrium state of the system is asymptotically stable in the large.

Solution In case (a), define the norm by

$$\|\mathbf{x}\| = \max_i \{c_i|x_i|\}$$

then

$$\begin{aligned}\|\mathbf{H}(\mathbf{x})\mathbf{x}\| &= \max_i \{c_i|\sum_{j=1}^n h_{ij}(\mathbf{x})x_j|\} \leqslant \max_i \left\{\sum_{j=1}^n \frac{c_i}{c_j}\left|h_{ij}(\mathbf{x})\right| \cdot c_j|x_j|\right\} \\ &\leqslant \max_i \left\{\sum_{j=1}^n \frac{c_i}{c_j}\left|h_{ij}(\mathbf{x})\right|\right\} \cdot \max_j \{c_j|x_j|\} < \max_j \{c_j|x_j|\} \\ &= \|\mathbf{x}\|\end{aligned}$$

which verifies that $\mathbf{H}(\mathbf{x})\mathbf{x}$ is a contraction.

In case (b), define the norm by

$$\|\mathbf{x}\| = \sum_{i=1}^n c_i|x_i|$$

Then

$$\|\mathbf{H}(\mathbf{x})\mathbf{x}\| = \sum_{i=1}^{n} c_i \left| \sum_{j=1}^{n} h_{ij}(\mathbf{x}) x_j \right| \leqslant \sum_{i=1}^{n} \frac{c_i}{c_j} \sum_{j=1}^{n} |h_{ij}(\mathbf{x})| \cdot c_j |x_j|$$

$$\leqslant \max_j \left\{ \sum_{i=1}^{n} \frac{c_i}{c_j} |h_{ij}(\mathbf{x})| \right\} \cdot \sum_{j=1}^{n} c_j |x_j| < \sum_{j=1}^{n} c_j |x_j|$$

$$= \|\mathbf{x}\|$$

which yields the desired result.

Problem A-8-30 Consider a discrete-time system

$$\mathbf{x}((k+1)T) = \mathbf{G}\mathbf{x}(kT)$$

where $\mathbf{x}$ is an n vector and $\mathbf{G}$ is an $n \times n$ constant matrix. Show that the scalar function $\mathbf{x}^*\mathbf{P}\mathbf{x}$ where $\mathbf{P}$ is a positive definite Hermitian matrix (or a positive definite real symmetric matrix) is a Liapunov function, if $\mathbf{P} - \mathbf{G}^*\mathbf{P}\mathbf{G}$ is positive definite.

Solution Let

$$V(\mathbf{x}) = \mathbf{x}^*\mathbf{P}\mathbf{x}$$

Then $V(\mathbf{x})$ is positive definite and $\Delta V(\mathbf{x}(kT))$ is given by

$$\Delta V(\mathbf{x}(kT)) = \mathbf{x}^*(kT)\mathbf{G}^*\mathbf{P}\mathbf{G}\mathbf{x}(kT) - \mathbf{x}^*(kT)\mathbf{P}\mathbf{x}(kT)$$

$$= -\mathbf{x}^*(kT)\mathbf{Q}\mathbf{x}(kT)$$

where $\mathbf{Q} = \mathbf{P} - \mathbf{G}^*\mathbf{P}\mathbf{G}$ = Hermitian (or real symmetric) matrix

Hence if $\mathbf{Q}$ is positive definite,

$$\Delta V < 0 \qquad \text{for all } \mathbf{x} \neq \mathbf{0}$$

Thus, $V(\mathbf{x}) = \mathbf{x}^*\mathbf{P}\mathbf{x}$ is a Liapunov function if $\mathbf{P} - \mathbf{G}^*\mathbf{P}\mathbf{G}$ is positive definite.

Problem A-8-31 Consider the following system:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}, \qquad \mathbf{x}(0) = \mathbf{c}$$

where $\mathbf{x}$ is an n vector and all the eigenvalues of $\mathbf{A}$ have negative real parts. Evaluate the following performance index:

$$J = \int_0^\infty \mathbf{x}^*\mathbf{Q}\mathbf{x}\,dt$$

where $\mathbf{Q}$ is a constant positive definite (or positive semidefinite) Hermitian or real symmetric matrix.

Suppose that the system equation is given by

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -2 & -a \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

and the performance index is given by

$$J = \int_0^\infty \mathbf{x}'\mathbf{x}\,dt$$

Determine the value of the parameter a (where $a > 0$) so as to minimize the performance index J when the initial condition is given by

$$\mathbf{x}(0) = \mathbf{c} = \begin{bmatrix} x_1(0) \\ 0 \\ 0 \end{bmatrix}$$

Solution Let

$$\mathbf{x}^*\mathbf{Q}\mathbf{x} = -\frac{d}{dt}(\mathbf{x}^*\mathbf{P}\mathbf{x})$$

where $\mathbf{P}$ is a constant positive definite Hermitian or real symmetric matrix. Then we obtain

$$\mathbf{x}^*\mathbf{Q}\mathbf{x} = -\dot{\mathbf{x}}^*\mathbf{P}\mathbf{x} - \mathbf{x}^*\mathbf{P}\dot{\mathbf{x}} = -\mathbf{x}^*\mathbf{A}^*\mathbf{P}\mathbf{x} - \mathbf{x}^*\mathbf{P}\mathbf{A}\mathbf{x}$$

Hence

$$\mathbf{A}^*\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{Q}$$

By the second method of Liapunov we know that if $\mathbf{A}$ is stable, then for a given $\mathbf{Q}$ there exists a $\mathbf{P}$ that satisfies this equation. Since the eigenvalues of $\mathbf{A}$ have negative real parts, $\mathbf{A}$ is stable and we can determine the elements of $\mathbf{P}$. J can be evaluated as

$$J = \int_0^\infty \mathbf{x}^*\mathbf{Q}\mathbf{x}\,dt = -\mathbf{x}^*\mathbf{P}\mathbf{x}\Big|_0^\infty$$

$$= -\mathbf{x}^*(\infty)\mathbf{P}\mathbf{x}(\infty) + \mathbf{x}^*(0)\mathbf{P}\mathbf{x}(0)$$

Since all the eigenvalues of $\mathbf{A}$ have negative real parts, $\mathbf{x}(\infty) \to \mathbf{0}$. Therefore, we obtain

$$J = \mathbf{x}^*(0)\mathbf{P}\mathbf{x}(0) = \mathbf{c}^*\mathbf{P}\mathbf{c}$$

To determine the parameter a in $\mathbf{A}$ of the third order system, we first determine $\mathbf{P}$ form

$$\mathbf{A}'\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{I}$$

Noting that $\mathbf{P}$ is symmetric, we have

$$\begin{bmatrix} 0 & 0 & -1 \\ 1 & 0 & -2 \\ 0 & 1 & -a \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{12} & p_{22} & p_{23} \\ p_{13} & p_{23} & p_{33} \end{bmatrix} + \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{12} & p_{22} & p_{23} \\ p_{13} & p_{23} & p_{33} \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -2 & -a \end{bmatrix}$$

$$= \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

Notice, however, that

$$J = \mathbf{c}'\mathbf{P}\mathbf{c}$$

$$= [x_1(0) \quad 0 \quad 0] \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{12} & p_{22} & p_{23} \\ p_{13} & p_{23} & p_{33} \end{bmatrix} \begin{bmatrix} x_1(0) \\ 0 \\ 0 \end{bmatrix}$$

$$= p_{11}x_1^2(0)$$

Hence we need know only p_{11}. By a simple computation p_{11} is found as

$$p_{11} = \frac{a^2 + 5a - 1}{4a - 2}$$

Therefore we obtain

$$J = \frac{a^2 + 5a - 1}{4a - 2} x_1^2(0)$$

To minimize J we determine a from

$$\frac{dJ}{da} = 0 \qquad \text{or} \qquad \frac{4a^2 - 4a - 6}{(4a - 2)^2} = 0$$

from which

$$a = 1.823 \qquad \text{or} \qquad -0.823$$

Notice that $d^2J/da^2 > 0$ if $a = 1.823$. Hence the optimal value of a (where $a > 0$) that minimizes J is 1.823.

The eigenvalues of $\mathbf{A}$ designed in this way have negative real parts because the designed system satisfies Liapunov's theorem for asymptotic stability. To confirm this let us use Routh's stability criterion. $\mathbf{A}$ can be written as

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -2 & -1.823 \end{bmatrix}$$

Hence the characteristic equation becomes

$$|\mathbf{A} - \lambda \mathbf{I}| = -\lambda^3 - 1.823\lambda^2 - 2\lambda - 1 = 0$$

By constructing the Routh array of coefficients

λ^3	1	2
λ^2	1.823	1
λ^1	1.45	0
λ^0	1	

we confirm that the eigenvalues of $\mathbf{A}$ have negative real parts.

PROBLEMS

Problem B-8-1 Consider the following equation:

$$\lambda^4 + 13\lambda^3 + 59\lambda^2 + 107\lambda + 60 = 0$$

Determine the number of roots whose real parts are greater than -1.

Problem B-8-2 Suppose that X and Y are functions of x and y. Prove that if there exists a functional relationship between X and Y, then

$$\frac{\partial(X, Y)}{\partial(x, y)} = 0$$

Problem B-8-3 Construct a Liapunov function for the following linear time-invariant system:

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = -6x_1 - 5x_2$$

Problem B-8-4 Consider the system

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = -\omega_n^2 x_1 + 2\zeta\omega_n x_2$$

where $\zeta > 0$ and $\omega_n > 0$. By use of the second method of Liapunov prove that the origin of the system is unstable.

Problem B-8-5 Consider a Hermitian matrix $\mathbf{A}$ and the characteristic polynomial

$$|\mathbf{A} - \lambda \mathbf{I}| = (-)^n[\lambda^n - b_1\lambda^{n-1} + b_2\lambda^{n-2} + \cdots + (-)^n b_n]$$

Prove that a necessary and sufficient condition that $\mathbf{A}$ is positive definite is that all the b_i $(i = 1, 2, \ldots, n)$ be positive.

Problem B-8-6 Consider the system

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(\mathbf{x}, t) \tag{8-63}$$

where $\mathbf{x}$ is an n vector and $\mathbf{A}(t)$ is continuous and bounded for $t \geqslant t_0$ and $\mathbf{B}(\mathbf{x}, t)$ is continuous for $\mathbf{x}$ near the origin and for $t \geqslant t_0$, and

$$\frac{\|\mathbf{B}(\mathbf{x}, t)\|}{\|\mathbf{x}\|} \to 0$$

uniformly as $\|\mathbf{x}\| \to 0$ for $t \geqslant t_0$. If all solutions of the following equation

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$$

approach the origin as $t \to \infty$, the same is true for all solutions of Eq. (8-63).

Problem B-8-7 Consider the following system:

$$\ddot{x} + a(x^2 + 1)\dot{x} + (x + bx^3) = 0$$
$$a > 0, \qquad b > 0$$

Show that the origin is asymptotically stable in the large.

Problem B-8-8 Consider the system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}), \qquad \mathbf{f}(\mathbf{0}) = \mathbf{0}$$

Prove that if the eigenvalues of the Jacobian matrix $\mathbf{J}(\mathbf{0})$, where

$$\mathbf{J}(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

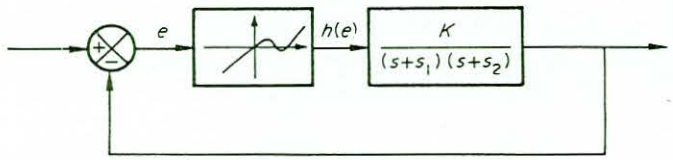
have negative real parts, then each solution which initiates near the origin must approach the origin as $t \to \infty$.

Problem B-8-9 Determine the stability of the origin of the following system:

$$\dot{x}_1 = -x_1 + x_2 - x_1(x_1^2 + x_2^2)$$

$$\dot{x}_2 = -x_1 - x_2 - x_2(x_1^2 + x_2^2)$$

Problem B-8-10 Determine the stability of the origin of the system shown in Fig. 8–9.



$$\int_0^e h(e)\,de \geq 0 \text{ for all } e \neq 0, \quad h(0) = 0$$

$$s_1 > 0, \qquad s_2 > 0$$

Figure 8-9 Nonlinear control system.

Problem B-8-11 (*8–2*) Consider the discrete-time system

$$\mathbf{x}((k+1)T) = \mathbf{G}\mathbf{x}(kT)$$

where $\mathbf{x}$ is an n vector and $\mathbf{G}$ is an $n \times n$ constant matrix. Prove that the eigenvalues of a constant matrix $\mathbf{G}$ are less than ρ in absolute value if and only if, given any positive definite real symmetric matrix $\mathbf{Q}$ the linear equation

$$\rho^{-2}\mathbf{G}'\mathbf{P}\mathbf{G} - \mathbf{P} = -\mathbf{Q}$$

has a unique, positive definite real symmetric solution $\mathbf{P}$.

Problem B-8-12 Consider the discrete-time system

$$\mathbf{x}((k+1)T) = \mathbf{G}\mathbf{x}(kT)$$

which is obtained by discretization from the linear continous-time system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$$

If the equilibrium state of the continuous-time system is asymptotically stable, then show that the equilibrium state of the discrete-time system is also asymptotically stable.

Problem B-8-13 Determine the stability of the equilibrium state of the following system:

$$x_1(k+1) = x_1(k) + ax_2(k)$$

$$x_2(k+1) = -ax_1(k) - (a-1)x_2(k) - ax_1(k-1)$$

where k denotes the kth sampling time and a is a constant.

Problem B-8-14 By use of a specific example show that although the output of a discrete-time system at sampling instants may be bounded, it does not imply that the output for $t_0 < t < \infty$ is bounded. (Note that since the effect of unstable poles can be canceled by the sampling operation, the output at sampling instants can be finite but growing oscillations may exist between sampling instants.)

Problem B-8-15 Prove that the maximum (minimum) value of the Hermitian form $\mathbf{x}^*\mathbf{A}\mathbf{x}$, subject to the condition that $\mathbf{x}^*\mathbf{x} = 1$, is equal to the greatest (least) eigenvalue of $\mathbf{A}$.

REFERENCES

8-1 Cesari, Lamberto, "Problems of Asymptotic Behavior and Stability," *AIEE CP* 60–1020.

8-2 Kalman, R. E., and J. E. Bertram, "Control System Analysis and Design Via the Second Method of Liapunov: I. Continuous-time Systems; II. Discrete-time Systems," *ASME J. Basic Engineering*, ser. **D**, **82** (1960), 371–93; 394–400.

8-3 LaSalle, J. P., "Some Extensions of Liapunov's Second Method," *RIAS Tech. Report* 60–5.

8-4 Cunningham, W. J., "An Introduction to Lyapunov's Second Method," in *Work Session in Lyapunov's Second Method*, ed. L. F. Kazda, September, 1960, University of Michigan.

8-5 Parks, P. C., "A New Proof of the Hurwitz Stability Criterion by the Second Method of Liapunov, with Applications to Optimum Transfer Functions," *JACC* Preprint (1963), 471–78.

8-6 Hahn, Wolfgang, *Theory and Application of Liapunov's Direct Method*, Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1963.

8-7 Lur'e, A. I., and E. N. Rozenvasser, "On Methods of Constructing Liapunov Functions in the Theory of Nonlinear Control Systems," *Proc. First Intern. Cong. IFAC*, Moscow, 1960, *Automatic and Remote Control*, London: Butterworths & Company, Ltd., 1961, pp. 928–33.

8-8 LaSalle, J. P., and S. Lefschetz, *Stability by Liapunov's Direct Method with Applications*, New York: Academic Press, 1961.

8-9 Gibson, J. E., and Z. V. Rekasius, "Application of Liapunov's Second Method to Control Systems with Nonlinear Gain," in *Work Session in Lyapunov's Second Method* ed. L. F. Kazda, September, 1960, University of Michigan.

8-10 Schultz, D. G., and J. E. Gibson, "The Variable Gradient Method for Generating Liapunov Functions," *AIEE Trans.* Part II, **81** (1962), 203–09.

8-11 Ingwerson, D. R., "A Modified Liapunov Method for Nonlinear Stability Analysis," *IRE Trans. Automatic Control* (1961), 199–210.

8-12 Gibson, J. E., *Nonlinear Automatic Control*, New York: McGraw-Hill Book Company, Inc., 1963.

8-13 Bellman, R., *Stability Theory of Differential Equations*, New York: McGraw-Hill Book Company, Inc., 1953.

8-14 Aizerman, M. A., *Theory of Automatic Regulation*, Moscow: Gostekhizdat, 1952.

8-15 Aizerman, M. A., and F. R. Gantmacher, "Conditions for the Existence of a Stability Domain for Single Looped Control Systems," *Prikladnaia Mat. i Mech.*, **18** (1954), 103–22.

8-16 Aizerman, M. A., "Conditions for Structural Stability of Certain Classes of Control Systems," *Regelungstechnik, Moderne Theorien und ihre Verwendbarkeit, Proc. Heidelberg Symposium*, September, 1956, Muenchen: Verlag R. Oldenbourg, 1957, pp. 93–98.

8-17 Lehnigk, S. H., "Die Strukturelle Stabilität Linearer Einläufiger PID-Regelkreise mit Differenzierender Wirkung Höchstens Zweiter Ordnung," *Regelungstechnik*, **6** (1958), 246–50.

8-18 Lehnigk, S. H., "On the Structural Stability of a Certain Class of Linear Motions," AGARD *Report* No. 194, 1958.

8-19 Lehnigk, S. H., "Die Strukturelle Stabilität der Bewegung von Flugzeugen," *Z. Flugw.*, **6** (1958), 110–15.

8-20 Margolis, S. G., and W. G. Vogt, "Control Engineering Applications of V. I. Zubov's Construction Procedure for Lyapunov Functions," *IRE Trans. Automatic Control* (1963), 104–13.

8-21 Persidkii, S. K., "On the Second Method of Liapunov," *J. Appl. Math. Mech.*, **25**, No. 1 (1961), 20–28.

8-22 Szegö, G. P., "Method of Constructing Liapunov Functions for Linear Non-Stationary Systems," *J. Elect. Cont.*, **13**, No. 1 (1962), 69–88.

8-23 Ergen, W. K., H. J. Lipkin, and J. A. Nohel, "Applications of Liapunov's Second Method in Reactor Dynamics," *J. Math. Phys.*, **36** (1957) 36–48.

8-24 Szegö, G. P., "On the Application of the Zubov Method for Construction of Liapunov's Functions for Nonlinear Control Systems," *JACC* Preprint (1962), Paper No. 2-2.

9

Optimal Control Systems

9-1 INTRODUCTION

Recently a number of techniques for designing optimal regulators and controllers have been proposed in the literature. In this chapter a direct method and a method using the functional equations technique of dynamic programming are presented for analytical design of optimal control systems. The plants to be considered in this chapter are linear and stationary and the performance indexes considered are quadratic or Hermitian. All control processes considered in this chapter are of finite duration and deterministic. The optimal control problem here is to find the control vector $\mathbf{u}(t)$ such as to have the state vector $\mathbf{x}(t)$ as close as possible to some ideal; for example, in the regulator problem to have $\mathbf{x}(t)$ as close to some constant vector as possible.

In practical optimization problems we must consider certain constraints on the allowable value of the control vector $\mathbf{u}(t)$. In some cases the control vector is subjected to saturation. In other cases the energy available for control may be limited. In this chapter optimal control systems are designed with constraints on the energy available for control. Mathematically the constraint on the available energy can be included in the performance indexes.

Consider a control system which is represented by the following state space equation:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \qquad \mathbf{x}(0) = \mathbf{c} \tag{9-1}$$

where

$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{A} = n \times n$ constant matrix
$\mathbf{B} = n \times r$ constant matrix

Equation (9–1) gives the relationship between the state vector $\mathbf{x}$ and the control vector $\mathbf{u}$. The problem is to steer the state vector $\mathbf{x}$ from an initial state $\mathbf{c}$ to some final state or region in the time interval $0 \leqslant t \leqslant T$. In doing this we wish to minimize the performance index which is given by

$$J(\mathbf{c}, T) = J_1(\mathbf{x}(T)) + \int_0^T J_2(\mathbf{x}, \mathbf{u}, t)\, dt \tag{9–2}$$

where $J_1(\mathbf{x}(T))$ is a function of $\mathbf{x}(T)$ and $J_2(\mathbf{x}, \mathbf{u}, t)$ is a function of $\mathbf{x}$, $\mathbf{u}$, and t, by choosing the optimal control vector $\mathbf{u}(t)$ in the interval $0 \leqslant t \leqslant T$. [Specific expressions for $J_1(\mathbf{x}(T))$ and $J_2(\mathbf{x}, \mathbf{u}, t)$ are given later in Sections 9–2 to 9–5.] Notice that the minimum value of the performance index is a function only of the initial state $\mathbf{c}$, the time duration T, and the control vector $\mathbf{u}$. In Eq. (9–2), J_1 is evaluated at the terminal state $\mathbf{x}(T)$ and J_2 is evaluated over the actual trajectory. In this chapter we limit these functions to be quadratic or Hermitian.

In general, different performance indexes lead to systems with different types of behavior. What kind of performance index leads to the most desirable system is largely a subjective question and cannot be answered without reference to a specific application.

It is a well known fact that the problem of minimizing the performance index for the general case is usually difficult to carry out and leads to a nonlinear feedback system whenever the performance index is of neither quadratic nor of Hermitian form. Thus, for nonquadratic or non-Hermitian performance indexes the solution may have to be obtained by numerical techniques by use of a digital computer. The possibility of solving the problem analytically whenever the performance index is of quadratic or Hermitian is a major advantage of the quadratic or Hermitian performance indexes.

Note that optimal control systems to be designed in this chapter based on given performance indexes are optimal for all initial conditions if all state variables can be measured. Note also that although the plants considered in this chapter are linear time-invariant, it will be seen in the following sections that if the duration of time in the control process is limited the optimal controllers become time-varying. (Hence such optimal control systems are time-varying even if the differential equation of the plant is time-invariant.)

This chapter is organized as follows: In Section 9–2, we present an analytical technique for designing noise-free optimal regulators. For given performance indexes we obtain by the direct method the optimal control vectors and the minimum values of performance indexes. Section 9–3 presents the functional equations technique of dynamic programming. We apply this technique to the analytical design of noise-free optimal regulators. Section 9–4 discusses analytical techniques for designing minimum error regulators. Finally, Section 9–5 presents analytical design of minimum error controllers with polynomial inputs.

9-2 ANALYTICAL DESIGN OF NOISE-FREE OPTIMAL REGULATORS

In guidance and orbit transfer problems, minimization of system terminal errors is important. Such a problem is usually called a *noise-free optimal regulator* problem. In this problem the system is initially at a state $\mathbf{c}$ which is different from the desired state. It is required to bring the system state as close to the desired terminal state, the origin of the state space, as possible by applying a control vector to the plant. The optimal design procedure can be applied to either a finite control interval $0 \leqslant t \leqslant T$ or infinite control interval $0 \leqslant t \leqslant \infty$. (This is in contrast to classical design methods which require that the control interval be infinity.) In this chapter we assume that the control interval is $0 \leqslant t \leqslant T$. In solving this problem we formulate it as that of a minimization problem, that is, the problem of minimizing, for example, the terminal error

$$\mathbf{x}^*(T)\mathbf{x}(T) = |x_1(T)|^2 + |x_2(T)|^2 + \cdots + |x_n(T)|^2$$

In this section we shall assume that the energy for control is specified, or

$$\int_0^T \mathbf{u}^*(t)\mathbf{u}(t)\, dt = \text{constant}$$

By use of a Lagrange multiplier λ, the constraint on the available energy can be included in the performance index. The performance index for this system can then be written as

$$J(\mathbf{c}, T) = \mathbf{x}^*(T)\mathbf{x}(T) + \lambda \int_0^T \mathbf{u}^*(t)\mathbf{u}(t)\, dt$$

The technique of the Lagrange multiplier is widely used for problems of this type, where one condition is to be a minimum under other restrictive conditions. In fact, the conditions need not be in integral forms, and any functional or differential relation among variables can be handled in a similar manner.

The problem here may be restated in a general form as follows: Given the plant governed by the following state space equation:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \qquad \mathbf{x}(0) = \mathbf{c} \tag{9–3}$$

where

$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{A} = n \times n$ constant matrix
$\mathbf{B} = n \times r$ constant matrix

and the performance index defined by

$$J(\mathbf{c}, T) = \mathbf{x}^*(T)\mathbf{P}\mathbf{x}(T) + \lambda \int_0^T \mathbf{u}^*(t)\mathbf{u}(t)\, dt$$

where $\mathbf{P} = (p_{ij})$ is a positive definite (or positive semidefinite) real symmetric or Hermitian matrix such that $p_{ii} = p_{jj}$ if x_i and x_j are a complex conjugate pair, and λ is a Lagrange multiplier which is a positive constant indicating the weight of control cost with respect to the minimizing errors, find the optimal control vector $\mathbf{u}(t)$ which minimizes the performance index. Notice that the elements of $\mathbf{P}$ depend on how the terminal errors are to be weighted. In this section we shall solve this problem by the direct method.

Note that in obtaining numerical solutions it is often advantageous to simplify the state space equation into a simpler one where the matrix $\mathbf{A}$ is in the Jordan canonical form. Notice that a quadratic performance index in terms of a real state vector may then become a Hermitian performance index with respect to a new state vector, if it is complex. Note also that if a state vector of a physical system is a complex vector, then any complex elements of the state vector always occur as complex conjugate pairs. Throughout this chapter equations for optimal control of linear systems are derived for the case involving complex vectors and complex matrices. Therefore, the notation of conjugate transpose appears frequently in equations. We shall, however, occasionally use the notation of transpose in particular examples if the system considered involves only real vectors and real matrices.

Solution of noise-free optimal regulator problem by the direct method *(9–1)*[1] The solution of Eq. (9–3) is given by

$$\mathbf{x}(t) = e^{\mathbf{A}t}\left[\mathbf{c} + \int_0^t e^{-\mathbf{A}s}\mathbf{B}\mathbf{u}(s)\, ds\right]$$

Let us denote by $J_0(\mathbf{c}, T)$ and $\mathbf{x}_0(t)$ the J and $\mathbf{x}$ corresponding to a special $\mathbf{u}_0$ where $\mathbf{u}_0$ will be specified later. Since

$$\mathbf{x}_0(t) = e^{\mathbf{A}t}\left[\mathbf{c} + \int_0^t e^{-\mathbf{A}s}\mathbf{B}\mathbf{u}_0(s)\, ds\right] \tag{9–4}$$

we have

$$\mathbf{x}(t) = \mathbf{x}_0(t) + \int_0^t e^{\mathbf{A}(t-s)}\mathbf{B}[\mathbf{u}(s) - \mathbf{u}_0(s)]\, ds$$

Then $J(\mathbf{c}, T)$ can be expressed as

$$J(\mathbf{c}, T) = J_0(\mathbf{c}, T) + \mathbf{x}_0^*(T)\mathbf{P}\int_0^T e^{\mathbf{A}(T-s)}\,\mathbf{B}[\mathbf{u}(s) - \mathbf{u}_0(s)]\, ds$$
$$+ \int_0^T [\mathbf{u}(s) - \mathbf{u}_0(s)]^*\mathbf{B}^*(e^{\mathbf{A}(T-s)})^* ds \mathbf{P}\mathbf{x}_0(T)$$

[1]Figures in parentheses refer to material listed at the end of the chapter.

$$+ \left[\int_0^T e^{\mathbf{A}(T-s)}\mathbf{B}[\mathbf{u}(s) - \mathbf{u}_0(s)]\, ds\right]^*\mathbf{P}\left[\int_0^T e^{\mathbf{A}(T-s)}\mathbf{B}[\mathbf{u}(s) - \mathbf{u}_0(s)]\, ds\right]$$
$$+ \lambda\int_0^T [\mathbf{u}(t) - \mathbf{u}_0(t)]^*[\mathbf{u}(t) - \mathbf{u}_0(t)]\, dt + \lambda\int_0^T \mathbf{u}_0^*(t)[\mathbf{u}(t) - \mathbf{u}_0(t)]\, dt$$
$$+ \lambda\int_0^T [\mathbf{u}(t) - \mathbf{u}_0(t)]^*\mathbf{u}_0(t)\, dt \tag{9–5}$$

We shall define $\mathbf{u}_0(t)$ by

$$\lambda\mathbf{u}_0(t) = -\mathbf{B}^*(e^{\mathbf{A}(T-t)})^*\mathbf{P}\mathbf{x}_0(T) \tag{9–6}$$

Then the second and the sixth terms on the right-hand side of Eq. (9–5) cancel each other. Similarly, the third and the last terms on the right-hand side of Eq. (9–5) cancel each other. Hence

$$J(\mathbf{c}, T) = J_0(\mathbf{c}, T) + \lambda\int_0^T [\mathbf{u}(t) - \mathbf{u}_0(t)]^*[\mathbf{u}(t) - \mathbf{u}_0(t)]\, dt$$
$$+ \left[\int_0^T e^{\mathbf{A}(T-s)}\mathbf{B}[\mathbf{u}(s) - \mathbf{u}_0(s)]\, ds\right]^*\mathbf{P}\left[\int_0^T e^{\mathbf{A}(T-s)}\mathbf{B}[\mathbf{u}(s) - \mathbf{u}_0(s)]\, ds\right] \tag{9–7}$$

Since $\lambda > 0$ and $\mathbf{P}$ is positive definite or positive semidefinite, Eq. (9–7) indicates that $\mathbf{u}_0$ gives the minimum value of $J(\mathbf{c}, T)$. Thus, $\mathbf{u}_0$ is the optimal control vector. From Eq. (9–4), we have

$$\mathbf{x}_0(T) = e^{\mathbf{A}T}\left[\mathbf{c} + \int_0^T e^{-\mathbf{A}s}\mathbf{B}\mathbf{u}_0(s)\, ds\right] \tag{9–8}$$

Substituting Eq. (9–6) into Eq. (9–8),

$$\mathbf{x}_0(T) = e^{\mathbf{A}T}\left[\mathbf{c} - \frac{1}{\lambda}\int_0^T e^{-\mathbf{A}s}\mathbf{B}\mathbf{B}^*(e^{\mathbf{A}(T-s)})^*\mathbf{P}\, ds\, \mathbf{x}_0(T)\right]$$

from which

$$\mathbf{x}_0(T) = \left[\mathbf{I} + \frac{1}{\lambda}\int_0^T e^{\mathbf{A}(T-s)}\mathbf{B}\mathbf{B}^*(e^{\mathbf{A}(T-s)})^*\mathbf{P}\, ds\right]^{-1} e^{\mathbf{A}T}\mathbf{c} \tag{9–9}$$

(The indicated inverse always exists.) Substituting Eq. (9-9) into Eq. (9-6),

$$\lambda\mathbf{u}_0(t) = -\mathbf{B}^*(e^{\mathbf{A}(T-t)})^*\mathbf{P}\left[\mathbf{I} + \frac{1}{\lambda}\int_0^T e^{\mathbf{A}(T-s)}\mathbf{B}\mathbf{B}^*(e^{\mathbf{A}(T-s)})^*\mathbf{P}\, ds\right]^{-1} e^{\mathbf{A}T}\mathbf{c}$$

$\mathbf{u}_0(t)$ can then be written as

$$\mathbf{u}_0(t) = -\mathbf{B}^*(e^{\mathbf{A}(T-t)})^*\mathbf{P}\left[\lambda\mathbf{I} + \int_0^T e^{\mathbf{A}(T-s)}\mathbf{B}\mathbf{B}^*(e^{\mathbf{A}(T-s)})^*\mathbf{P}\, ds\right]^{-1} e^{\mathbf{A}T}\mathbf{c}$$

or

$$\mathbf{u}_0(t) = -\mathbf{B}^*(e^{\mathbf{A}(T-t)})^*\mathbf{P}\left[\lambda\mathbf{I} + \int_0^T e^{\mathbf{A}\tau}\mathbf{B}\mathbf{B}^*(e^{\mathbf{A}\tau})^*\mathbf{P}\, d\tau\right]^{-1} e^{\mathbf{A}T}\mathbf{c} \tag{9–10}$$

In particular, we have at $t = 0$

$$\mathbf{u}_0(0) = -\mathbf{B}^*(e^{\mathbf{A}T})^*\mathbf{P}\left[\lambda\mathbf{I} + \int_0^T e^{\mathbf{A}s}\mathbf{B}\mathbf{B}^*(e^{\mathbf{A}s})^*\mathbf{P}\, ds\right]^{-1} e^{\mathbf{A}T}\mathbf{c} \tag{9–11}$$

Eq. (9–11) gives the optimal control vector at $t = 0$.

Note that Eq. (9–10) specifies the optimal control law $\mathbf{u}_0(t)$ as a function of the initial state of the plant. This means that Eq. (9–10) gives an open loop optimal control law. In controlling a plant by use of the open loop optimal control law, the control vector for the control duration is preprogrammed on the basis of the initial state of the plant, $\mathbf{x}(0) = \mathbf{c}$. Open loop optimal control is optimal only in the absence of the disturbances.

In general, the feedback optimal control law where the control vector is specified as a function of the present state of the plant is better than the open loop optimal control law. By means of feedback optimal control, we can utilize the most up-to-date information on the state of the plant. In the presence of external disturbances and plant parameter variations, optimal control or nearly optimal control of a plant is possible only by feedback optimal control.

Let us derive the feedback optimal control law from Eq. (9–10) or Eq. (9–11). This is possible by the following procedure: Let $T - t$ be the duration of time remaining in the process and $\mathbf{x}(t)$ be the current state. Equation (9–11) can then be modified into

$$\mathbf{u}_0(t) = -\mathbf{B}^*(e^{\mathbf{A}(T-t)})^*\mathbf{P}\left[\lambda\mathbf{I} + \int_0^{T-t} e^{\mathbf{A}s}\mathbf{B}\mathbf{B}^*(e^{\mathbf{A}s})^*\mathbf{P}\,ds\right]^{-1} e^{\mathbf{A}(T-t)}\mathbf{x}(t) \tag{9–12}$$

[Note that in the noise-free optimal control system, $\mathbf{x}(t) = \mathbf{x}_0(t)$.]

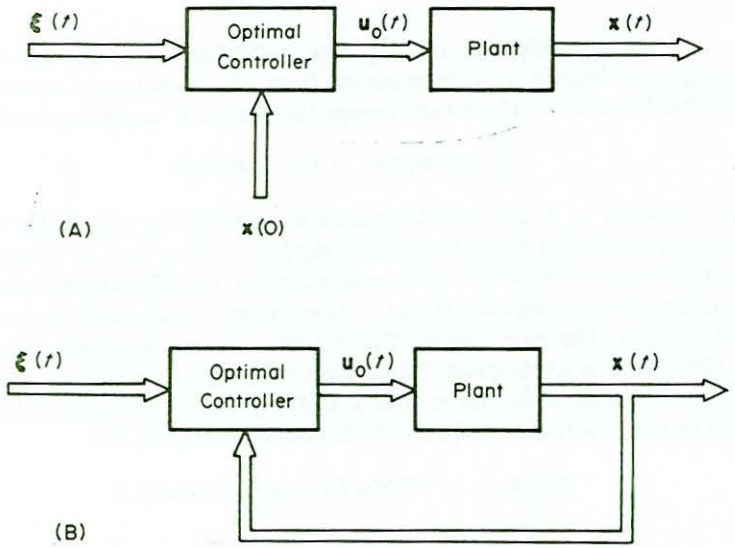


Figure 9-1 Block diagrams for optimal control systems. (A) Open loop optimal control system; (B) feedback optimal control system.

Equation (9–12) specifies the optimal control vector $\mathbf{u}_0(t)$ as a function of the present state $\mathbf{x}(t)$ and the remaining time in the process, $T - t$. Thus Eq. (9–12) gives the feedback optimal control law. Figures 9–1(A) and (B) illustrate an open loop optimal control system and a feedback optimal control system, respectively. [In the present optimal regulator problem the input $\xi(t)$ is zero for $0 \leqslant t \leqslant T$.]

Consider the control law defined by

$$\mathbf{u}(t) = -\mathbf{F}(T - t)\,\mathbf{x}(t)$$

where

$$\mathbf{F}(T - t) = \mathbf{B}^*(e^{\mathbf{A}(T-t)})^*\mathbf{P}\left[\lambda\mathbf{I} + \int_0^{T-t} e^{\mathbf{A}s}\mathbf{B}\mathbf{B}^*(e^{\mathbf{A}s})^*\mathbf{P}\,ds\right]^{-1} e^{\mathbf{A}(T-t)} \tag{9–13}$$

Clearly, this control law is the optimal control law for the present regulator problem. Notice that the explicit form of the optimal control law is determined by the matrix $\mathbf{F}(T - t)$. The block diagram of the control system based on this law, the feedback optimal control law, is shown in Fig. 9–2. The optimal control vector is a linear function of $\mathbf{x}(t)$, because the performance index is of quadratic or Hermitian form. The time-varying element $\mathbf{F}(T - t)$, the optimal regulator, can be considered as time-varying gains. If the measurements of the state variables and also $T - t$, the time remaining in the process, are readily available and reasonably accurate, the optimal control system thus designed will exhibit high performance.

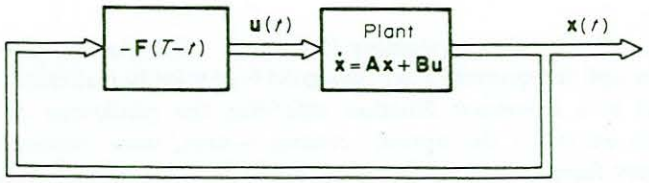


Figure 9-2 Optimal control system.

One elementary implementation of the time-varying gain is as follows: If the remaining time $T - t$ is available as an angular shaft position, the time-varying multiplications can be realized from a set of loaded potentiometers. The signal on the wiper of such a potentiometer would then be the appropriate function of the shaft angle $T - t$ times the excitation voltage applied to the loaded pot. The shaft angle $T - t$ is achieved by placing an input signal $T - t$ to a position servo.

Next, we shall obtain $J_0(\mathbf{c}, T)$, the minimum value of $J(\mathbf{c}, T)$. Define

$$f(\mathbf{c}, T) = J_0(\mathbf{c}, T)$$

Then by use of Eq. (9–6) and Eq. (9–9), we find

$$f(\mathbf{c}, T) = J_0(\mathbf{c}, T) = \mathbf{x}_0^*(T)\mathbf{P}\mathbf{x}_0(T) + \lambda \int_0^T \mathbf{u}_0^*(t)\mathbf{u}_0(t)\, dt$$

$$= \mathbf{x}_0^*(T)\mathbf{P}\mathbf{x}_0(T) + \frac{1}{\lambda}\mathbf{x}_0^*(T)\mathbf{P}\int_0^T e^{\mathbf{A}(T-t)}\mathbf{B}\mathbf{B}^*(e^{\mathbf{A}(T-t)})^*\mathbf{P}\, dt\, \mathbf{x}_0(T)$$

$$= \mathbf{x}_0^*(T)\mathbf{P}\left[\mathbf{I} + \frac{1}{\lambda}\int_0^T e^{\mathbf{A}(T-t)}\mathbf{B}\mathbf{B}^*(e^{\mathbf{A}(T-t)})^*\mathbf{P}\, dt\right]\mathbf{x}_0(T)$$

$$= \mathbf{c}^*(e^{\mathbf{A}T})^*\left[\mathbf{I} + \frac{1}{\lambda}\mathbf{P}\int_0^T e^{\mathbf{A}s}\mathbf{B}\mathbf{B}^*(e^{\mathbf{A}s})^*\, ds\right]^{-1}\mathbf{P}e^{\mathbf{A}T}\mathbf{c}$$

$$= \mathbf{c}^*\left\{\lambda(e^{\mathbf{A}T})^*\mathbf{P}\left[\lambda\mathbf{I} + \int_0^T e^{\mathbf{A}s}\mathbf{B}\mathbf{B}^*(e^{\mathbf{A}s})^*\mathbf{P}\, ds\right]^{-1}e^{\mathbf{A}T}\right\}\mathbf{c} \qquad (9\text{–}14)$$

Thus, $f(\mathbf{c}, T)$ is found to be a quadratic or Hermitian form in $\mathbf{c}$. In terms of the remaining time $T - t$ and the current state $\mathbf{x}(t)$, the minimized performance index can be given by

$$f(\mathbf{x}(t), T - t)$$

$$= \mathbf{x}^*(t)\left\{\lambda(e^{\mathbf{A}(T-t)})^*\mathbf{P}\left[\lambda\mathbf{I} + \int_0^{T-t} e^{\mathbf{A}s}\mathbf{B}\mathbf{B}^*(e^{\mathbf{A}s})^*\mathbf{P}\, ds\right]^{-1}e^{\mathbf{A}(T-t)}\right\}\mathbf{x}(t) \qquad (9\text{–}15)$$

The optimal control law, Eq. (9–12), previously obtained is based on the assumption that the exact information on the present state of the plant is available. In practice, however, this may not be the case. If disturbances are present, we must treat the problem from a stochastic viewpoint. If the state variables of the plant are not all accessible for direct measurement, optimal control of the plant must be accomplished by use of the best approximation $\hat{\mathbf{x}}(t)$ of the actual state $\mathbf{x}(t)$. The result in this case is, of course, not optimal but is nearly optimal.

Question of stability *(9–2, 9–3)* It is necessary to make sure that the optimal control system designed is physically realizable. This can be assured if a Liapunov function satisfying the conditions of asymptotic stability exists for the optimal control system, since existence of such a Liapunov function means that the equilibrium state of the system is asymptotically stable and the system can be physically realized.

In the present problem, if $\mathbf{P}$ is positive definite, then $f(\mathbf{c}, T)$ is a Liapunov function. This can be seen as follows: From Eq. (9–14) the expression for $f(\mathbf{c}, T)$ is always positive if $\mathbf{c} \neq \mathbf{0}$ and is zero only if $\mathbf{c} = \mathbf{0}$. Hence $f(\mathbf{c}, T)$ is positive definite. That df/dt is a nonpositive function of time may be seen from the physical meaning of $f(\mathbf{c}, T)$. From Eq. (9–14) we see that, once the initial condition $\mathbf{c}$ is given, $f(\mathbf{c}, T)$ is fixed. Since the control energy spent along the trajectory for any time interval $t_1 \leqslant t \leqslant t_2$, or

$$\int_{t_1}^{t_2} \mathbf{u}_0^*(t)\mathbf{u}_0(t)\, dt$$

is always a positive quantity, the function $f(\mathbf{c}, T)$ along the trajectory always decreases with time, or

$$\frac{df}{dt} < 0 \qquad \text{if } \mathbf{c} \neq \mathbf{0}$$

df/dt becomes zero only if $\mathbf{c} = \mathbf{0}$. Hence df/dt is negative definite. We thus see that $f(\mathbf{c}, T)$ is a Liapunov function satisfying the conditions for asymptotic stability. This is true for any $T > 0$ including $T = \infty$. We can now conclude that the optimal control system designed in this way is always asymptotically stable for $0 \leqslant t \leqslant T$.

So far we have been concerned with the system behavior for the time interval $0 \leqslant t \leqslant T$ and nothing is said about the system behavior for $t > T$. The system behavior for $t > T$ is determined by the control vector $\mathbf{u}$ for $t > T$. If we simply use

$$\mathbf{u}(t) = -\mathbf{F}(0)\,\mathbf{x}(t)$$

for $t > T$, then the stability of the system for $t > T$ is determined by the eigenvalues of the coefficient matrix of the following system:

$$\dot{\mathbf{x}} = [\mathbf{A} - \mathbf{B}\mathbf{F}(0)]\mathbf{x}$$

If the eigenvalues of $\mathbf{A} - \mathbf{B}\mathbf{F}(0)$ possess all negative real parts, then the system is also asymptotically stable for $t > T$. Otherwise, a stability device which does not go into action until the final instant of the transient period is needed to avoid running away from the final state $\mathbf{x}(T)$ after the duration of the control process is over.

Determination of Lagrange multiplier λ The value of the Lagrange multiplier λ is determined from the constraint on the energy available for control. If the total energy for control is specified, namely

$$\int_0^T \mathbf{u}^*(t)\mathbf{u}(t)\, dt = K = \text{constant}$$

then the value of λ can be determined by substituting Eq. (9–10) into the preceding constraint equation and solving for λ. Clearly, λ is a function of the initial state $\mathbf{c}$. Note that as the amount of the available energy is decreased the value of λ is increased. (Hence, λ is sometimes interpreted as the "price" of the energy. The more limited the available energy, the higher the price.)

Note that in some cases the amount of energy available is not limited, but we wish to minimize the sum of $\mathbf{x}^*(T)\,\mathbf{P}\mathbf{x}(T)$ and the cost of total energy spent in the process; namely, minimize the following $J(\mathbf{c}, T)$:

$$J(\mathbf{c}, T) = \mathbf{x}^*(T)\mathbf{P}\mathbf{x}(T) + \mu\int_0^T \mathbf{u}^*(t)\mathbf{u}(t)\, dt$$

where μ is the known price of the energy per unit amount. Then, this problem can be solved in exactly the same way as the problem just solved which involves the Lagrange multiplier λ. In this chapter the optimal control vector is obtained in terms of λ, which is either given (in which case $\lambda = \mu =$ known quantity) or can be determined once the total energy available for control is specified (in which case λ is a Lagrange multiplier).

Example 9-1 Consider the control system shown in Fig. 9–3. The

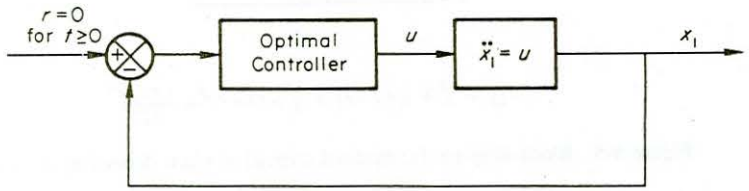


Figure 9-3 Control system.

input $r(t)$ to the system is removed at $t = 0$, or $r(t) = 0$ for $t \geqslant 0$. The performance index for this system is given by

$$J(\mathbf{c}, T) = \mathbf{x}'(T)\mathbf{x}(T) + \lambda \int_0^T u'(t)u(t)\, dt$$

where $\mathbf{x}$ and u are real and λ is a positive constant. The time interval $0 \leqslant t \leqslant T$ may be chosen any duration during which the command input to the system or external disturbances are zero and during which the plant to be controlled moves from one essential level of operation to another. It is desired to minimize the performance index. Obtain the optimal control function $u_0(t)$ and the minimized performance index $f(\mathbf{c}, T)$.

The system differential equation may be written in the form of the following state space equation:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u, \qquad \mathbf{x}(0) = \mathbf{c}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

From Eq. (9–12) the optimal control function $u_0(t)$ with $\mathbf{P} = \mathbf{I}$ is given as follows:

$$u_0(t) = -\mathbf{B}'(e^{\mathbf{A}(T-t)})'\left[\lambda\mathbf{I} + \int_0^{T-t} e^{\mathbf{A}s}\mathbf{B}\mathbf{B}'(e^{\mathbf{A}s})'\, ds\right]^{-1} e^{\mathbf{A}(T-t)}\mathbf{x}(t)$$

For this problem $e^{\mathbf{A}s}$ can be easily obtained as

$$e^{\mathbf{A}s} = \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix}$$

The integral expression becomes

$$\int_0^{T-t} e^{\mathbf{A}s}\mathbf{B}\mathbf{B}'(e^{\mathbf{A}s})'\, ds = \int_0^{T-t} \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix}[0 \quad 1]\begin{bmatrix} 1 & 0 \\ s & 1 \end{bmatrix} ds$$

$$= \begin{bmatrix} \dfrac{(T-t)^3}{3} & \dfrac{(T-t)^2}{2} \\ \dfrac{(T-t)^2}{2} & T-t \end{bmatrix}$$

Thus

$$\left[\lambda\mathbf{I} + \int_0^{T-t} e^{\mathbf{A}s}\mathbf{B}\mathbf{B}'(e^{\mathbf{A}s})'\, ds\right]^{-1}$$

$$= \frac{1}{\Delta}\begin{bmatrix} \lambda + (T-t) & -\dfrac{(T-t)^2}{2} \\ -\dfrac{(T-t)^2}{2} & \lambda + \dfrac{(T-t)^3}{3} \end{bmatrix} \equiv \mathbf{Z}^{-1}$$

where

$$\Delta = \lambda^2 + \lambda(T - t) + \frac{1}{3}\lambda(T-t)^3 + \frac{(T-t)^4}{12}$$

Hence

$$u_0(t) = -\mathbf{B}'(e^{\mathbf{A}(T-t)})'\mathbf{Z}^{-1}e^{\mathbf{A}(T-t)}\mathbf{x}(t)$$

$$= -\frac{\{\lambda(T-t) + [(T-t)^2/2]\}x_1(t) + \{\lambda + \lambda(T-t)^2 + [(T-t)^3/3]\}x_2(t)}{\lambda^2 + \lambda(T-t) + \lambda[(T-t)^3/3] + [(T-t)^4/12]}$$

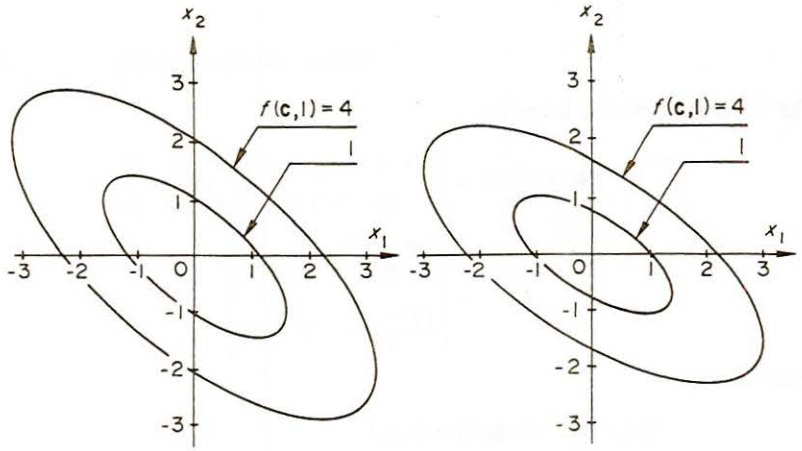
The minimized performance index $f(\mathbf{c}, T)$ is given by

$$f(\mathbf{c}, T) = \mathbf{c}^*\left\{\lambda(e^{\mathbf{A}T})'\left[\lambda\mathbf{I} + \int_0^T e^{\mathbf{A}s}\mathbf{B}\mathbf{B}'(e^{\mathbf{A}s})'\, ds\right]^{-1} e^{\mathbf{A}T}\right\}\mathbf{c}$$

$$= \frac{\lambda\{(\lambda + T)c_1^2 + 2[\lambda T + (T^2/2)]c_1c_2 + [\lambda + \lambda T^2 + (T^3/3)]c_2^2\}}{\lambda^2 + \lambda T + \lambda(T^3/3) + (T^4/12)}$$

Figure 9–4 shows families of curves $f(\mathbf{c}, T)$ with three different combinations of T and λ. It is seen in Fig. 9–4 that, for a given T and a given value of $f(\mathbf{c}, T)$, increasing the value of λ from 1 to 5 decreases the area enclosed by the $f(\mathbf{c}, T)$ curve. This indicates that the allowable initial magnitude of the error state is decreased. Hence, the effect of increasing λ for a constant T is to increase the terminal error. This may seem rather obvious because, as λ is increased, the terminal error is penalized less severely since the constraint on the energy for control is weighted more heavily. The effect on the terminal error of increasing T for a constant λ is a decrease in the terminal error. It is clearly seen in Fig. 9–4 that for a constant λ, areas enclosed by $f(\mathbf{c}, T)$ curves are increased as T is increased.

Figure 9–5 is a block diagram of the optimal control system. The system involves time-varying gains. The block diagram of Fig. 9–6 shows a conventional control of



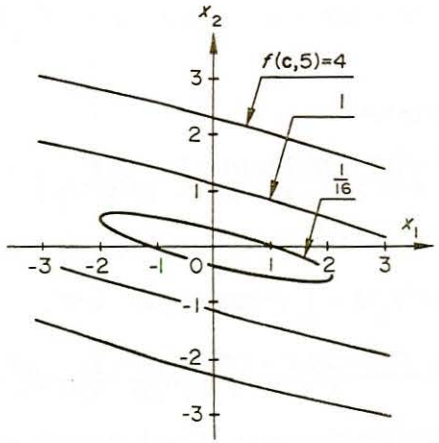
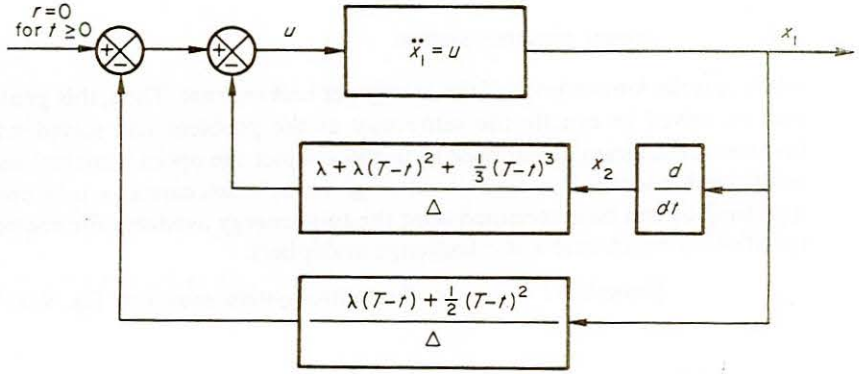


Figure 9-4 Families of curves $f(c, T)$.



$$\Delta = \lambda^2 + \lambda(T-t) + \frac{1}{3}\lambda(T-t)^3 + \frac{(T-t)^4}{12}$$

Figure 9-5 Block diagram for optimal control of plant defined by $\ddot{x}_1 = u$.

the plant $1/s^2$. The control system designed by the conventional method is a linear time-invariant system and the control function u assumes the form $u = -ax_1 - bx_2$, where a and b are constants. The compensating element of the linear time-invariant system can be a simple RC circuit, whereas the optimal control system designed in

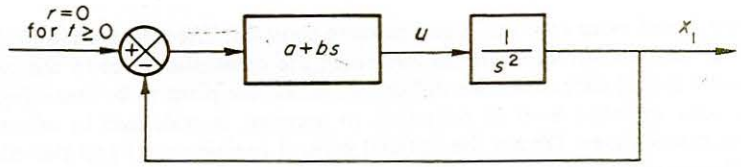


Figure 9-6 Block diagram for conventional control of plant defined by $1/s^2$.

this section is time-varying, and the essential control unit has time-varying gains which are much more complicated than an RC circuit. Although the control system designed by the conventional method may be satisfactory in performance, the control system designed here which involves time-varying gains is guaranteed to give the optimal performance under the present performance index.

Note, however, that a decision on whether to build the optimal controller or a simpler but inferior controller depends on a measure of the degree to which the performance of the optimal system, if realized, exceeds that of the simple system.

Finally, we shall examine stability of the optimal control system designed. The system designed is asymptotically stable during the time interval $0 \leqslant t \leqslant T$, the

duration of the control process, since the minimized performance index $f(\mathbf{c}, T)$ is a Liapunov function. After the end of the duration of the control process, if the following control law is used

$$u_0(t) = -\mathbf{F}(0)\mathbf{x}(t), \qquad t > T$$

then

$$u_0(t) = -\begin{bmatrix} 0 & \frac{1}{\lambda} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

Since the matrices $\mathbf{A}$ and $\mathbf{B}$ are given by

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

the matrix $\mathbf{A} - \mathbf{BF}(0)$ becomes as follows:

$$\mathbf{A} - \mathbf{BF}(0) = \begin{bmatrix} 0 & 1 \\ 0 & -\frac{1}{\lambda} \end{bmatrix}$$

The eigenvalues of $\mathbf{A} - \mathbf{BF}(0)$ are 0 and $-\frac{1}{\lambda}$. Hence the origin (the equilibrium state) of the system is not asymptotically stable for $t > T$. Therefore, we need some device preventing drifting away from the final state $\mathbf{x}(T)$ after the duration of the control process is over.

Example 9-2 Consider the control system shown in Fig. 9–7. This is an example of control of an unstable system. The input $r(t)$ to the system is removed at $t = 0$, or $r(t) = 0$ for $t \geqslant 0$. The performance index is given by

$$J(\mathbf{c}, T) = x'(T)x(T) + \lambda \int_0^T u'(t)u(t)\, dt$$

where x and u are real and λ is a positive constant. Find the optimal control function $u_0(t)$.

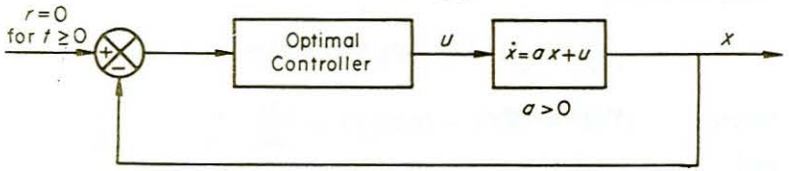


Figure 9-7 Control system.

The system differential equation is

$$\dot{x} = ax + u, \qquad a > 0, \qquad x(0) = c$$

Hence $\mathbf{A} = A = a$ and $\mathbf{B} = B = 1$. The optimal control function $u_0(t)$ is given by

$$u_0(t) = -B'(e^{A(T-t)})'\left[\lambda + \int_0^{T-t} e^{As} BB'(e^{As})'\, ds\right]^{-1} e^{A(T-t)}x(t)$$

$$= -\frac{2a}{1 + (2a\lambda - 1)e^{-2a(T-t)}}x(t), \qquad 0 \leqslant t \leqslant T$$

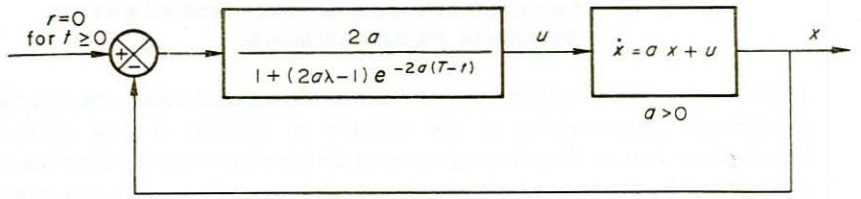


Figure 9-8 Block diagram for optimal control of plant defined by $\dot{x} = ax + u$.

Figure 9–8 shows configuration of the optimal control system. The minimized performance index $f(c, T)$ can be found as follows:

$$f(c, T) = c'\left\{\lambda(e^{AT})'\left[\lambda + \int_0^T e^{As}BB'(e^{As})'\, ds\right]^{-1} e^{AT}\right\}c$$

$$= \frac{2a\lambda}{1 + (2a\lambda - 1)e^{-2aT}}c^2$$

This is a Liapunov function. We can easily confirm this as follows: We may write $f(x(t), T - t)$ in the following form:

$$f(x(t), T - t) = \frac{2a\lambda}{1 + (2a\lambda - 1)e^{-2a(T-t)}}x(t)^2, \qquad 0 \leqslant t \leqslant T$$

Since $a > 0$, $\lambda > 0$, and $1 + (2a\lambda - 1)e^{-2a(T-t)} > 0$ we see that $f(x(t), T - t)$ is positive when $x(t) \neq 0$ and is zero only if $x(t) = 0$. Hence, it is positive definite. The time derivative of $f(x(t), T - t)$ is found as follows:

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial x}\dot{x}$$

$$= \frac{-4a^2\lambda}{[1 + (2a\lambda - 1)e^{-2a(T-t)}]^2}x(t)^2$$

df/dt is thus a nonpositive function of time and is zero only if $x(t) = 0$; df/dt is thus negative definite for $0 \leqslant t \leqslant T$. Thus, $f(c, T)$ is a Liapunov function for any positive value of λ. The system designed by the present method is asymptotically stable for $0 \leqslant t \leqslant T$ (where T can be infinity) and hence it is physically realizable. At $t = T$, the end of the control process, however, the system may run away from the final position $x(T)$ as time goes on unless the system designed is asymptotically stable for $t > T$.

If we assume that the time-varying term would remain constant at the value at $t = T$, namely,

$$u_0(t) = -F(0)x(t), \qquad t > T$$

then we obtain

$$u_0(t) = -\frac{1}{\lambda}x(t), \qquad t > T$$

The system differential equation for $t > T$ is then given by

$$\dot{x} = \left(a - \frac{1}{\lambda}\right)x$$

For $\lambda < 1/a$, the optimal control system is asymptotically stable and no separate stability device may be necessary at the end of the process, $t = T$. However, for $\lambda \geqslant 1/a$, a separate stability device is necessary in order to avoid running away from the final position $x(T)$.

9-3 FUNCTIONAL EQUATIONS TECHNIQUE OF DYNAMIC PROGRAMMING

This section presents application of the basic functional equations technique of dynamic programming to the solution of optimal control problems. We shall see that in determining optimal decision rules for continuous control systems by the functional equations technique of dynamic programming, we must solve partial differential equations which must be satisfied by the function representing the minimized performance index. The partial differential equation is derived from a recurrence relation obtained by applying the principle of optimality to the minimized performance index. The principle of optimality due to Bellman states[1] "An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision." The mathematical transliteration of this principle yields the functional equations we shall need in solving the minimization problem.

Consider the control system defined by Eq. (9–1), rewritten thus:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \qquad \mathbf{x}(0) = \mathbf{c}$$

where
$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{A} = n \times n$ constant matrix
$\mathbf{B} = n \times r$ constant matrix

and the performance index defined by Eq. (9–2), rewritten thus:

$$J(\mathbf{c}, T) = J_1(\mathbf{x}(T)) + \int_0^T J_2(\mathbf{x}, \mathbf{u}, t)\, dt$$

Our problem is minimizing the performance index $J(\mathbf{c}, T)$ by optimally selecting the control vector $\mathbf{u}(t)$ in the time interval $0 \leqslant t \leqslant T$. We clearly see that the problem here is that of the variational problem of minimizing $J(\mathbf{c}, T)$ subject to Eq. (9–1) with constraints on the control vector $\mathbf{u}$. By the functional equations technique of dynamic programming, we imbed this problem within the family of problems generated by allowing $\mathbf{c}$ and T to be parameters.

Let us define the minimum value of $J(\mathbf{c}, T)$ over all allowable $\mathbf{u}$ as $f(\mathbf{c}, T)$, namely

$$f(\mathbf{c}, T) = \min_{\mathbf{u}} J(\mathbf{c}, T) \tag{9–16}$$

[1]For detailed discussions of dynamic programming see (*9-4*, *9-5*).

We may interpret $f(\mathbf{c}, T)$ to represent a cost of a control process of duration T, the initial state being $\mathbf{c}$, and an optimal control policy being used. Let the optimal solution evolve from state $\mathbf{c}$ at time 0 to state $\mathbf{c} + \Delta\mathbf{c}$ at time Δt. At the new state we see that the same problem exists and, from Eq. (9–16), the variational scheme gives

$$\begin{aligned} f(\mathbf{c}, T) &= \min_{\mathbf{u}} \left[J_1 + \int_0^T J_2\, dt \right] \\ &= \min_{\mathbf{u}(0<t<\Delta t)} \min_{\mathbf{u}(\Delta t<t<T)} \left[\int_0^{\Delta t} J_2\, dt + J_1 + \int_{\Delta t}^T J_2\, dt \right] \\ &= \min_{\mathbf{u}} \left[\int_0^{\Delta t} J_2\, dt + f(\mathbf{c} + \Delta\mathbf{c}, T - \Delta t) \right] \end{aligned} \tag{9–17}$$

where $\mathbf{c} + \Delta\mathbf{c}$ is given by $\mathbf{x}(\Delta t)$, since

$$\mathbf{x}(\Delta t) = \mathbf{c} + \frac{\Delta\mathbf{c}}{\Delta t}\Delta t = \mathbf{c} + \Delta\mathbf{c}$$

The minimization is taken over all allowable control vectors $\mathbf{u}$ defined over $0 \leqslant t \leqslant T$ with $\mathbf{x}(0) = \mathbf{c}$. Equation (9–17) is the basic recurrence equation for the solution of the problem. Expanding Eq. (9–17) into a Taylor series assuming continuity of the partials,

$$f(\mathbf{c}, T) = \min_{\mathbf{u}} \left[J_2(\mathbf{x}, \mathbf{u}, t)\Delta t + f(\mathbf{c}, T) + \Delta c_1 \frac{\partial f}{\partial c_1} + \Delta c_2 \frac{\partial f}{\partial c_2} + \cdots + \Delta c_n \frac{\partial f}{\partial c_n} - \Delta t \frac{\partial f}{\partial T} + \cdots \right]$$

Canceling $f(\mathbf{c}, T)$ on both sides of the equation

$$\Delta t \frac{\partial f}{\partial T} = \min_{\mathbf{u}} \left[J_2(\mathbf{x}, \mathbf{u}, t)\Delta t + \left(\frac{\Delta c_1}{\Delta t}\frac{\partial f}{\partial c_1} + \frac{\Delta c_2}{\Delta t}\frac{\partial f}{\partial c_2} + \cdots + \frac{\Delta c_n}{\Delta t}\frac{\partial f}{\partial c_n} \right)\Delta t + 0[(\Delta t)^2] \right]$$

Dividing both sides of the equation by Δt and then letting Δt approach zero, we obtain

$$\frac{\partial f}{\partial T} = \min_{\mathbf{u}} \left[J_2(\mathbf{c}, \mathbf{u}(0), 0) + (\nabla h)^*[\mathbf{A}\mathbf{c} + \mathbf{B}\mathbf{u}(0)] \right] \tag{9–18}$$

where
$$(\nabla h)^* = (\nabla f)' = (\operatorname{grad} f)' = \begin{bmatrix} \frac{\partial f}{\partial c_1} & \frac{\partial f}{\partial c_2} & \cdots & \frac{\partial f}{\partial c_n} \end{bmatrix}$$

and

$$\mathbf{A}\mathbf{c} + \mathbf{B}\mathbf{u}(0) = \lim_{\Delta t \to 0} \begin{bmatrix} \frac{\Delta c_1}{\Delta t} \\ \vdots \\ \frac{\Delta c_n}{\Delta t} \end{bmatrix}$$

Equation (9–18) is the basic partial differential equation satisfied by the minimized performance index. The control vector $\mathbf{u}$ which minimizes the right-hand side of Eq. (9–18) is the optimal control vector. Equation (9–18) forms the basis of the operation of the optimal controller. It must be solved either analytically or numerically in order to determine the optimal control vector which minimizes the performance index.

Solution of noise-free optimal regulator problem by the functional equations technique of dynamic programming *(9–5, 9–6, 9–7)* Consider the plant governed by the following state space equation:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \qquad \mathbf{x}(0) = \mathbf{c} \tag{9–19}$$

where
$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{A} = n \times n$ constant matrix
$\mathbf{B} = n \times r$ constant matrix

and the performance index

$$J(\mathbf{c}, T) = \mathbf{x}^*(T)\mathbf{P}\mathbf{x}(T) + \lambda \int_0^T \mathbf{u}^*(t)\mathbf{u}(t)\, dt$$

where $\mathbf{P} = (p_{ij})$ is a positive definite (or positive semidefinite) real symmetric or Hermitian matrix such that $p_{ii} = p_{jj}$ if x_i and x_j are a complex conjugate pair, and λ is a Lagrange multiplier which is a positive constant. Our problem here is to find the optimal control vector $\mathbf{u}(t)$. Since the problem is to minimize $J(\mathbf{c}, T)$ over all allowable $\mathbf{u}$, define the minimum of the performance index as $f(\mathbf{c}, T)$, or

$$f(\mathbf{c}, T) = \min_{\mathbf{u}} \left[\mathbf{x}^*(T)\mathbf{P}\mathbf{x}(T) + \lambda \int_0^T \mathbf{u}^*(t)\mathbf{u}(t)\, dt \right]$$

$f(\mathbf{c}, T)$ can be rewritten as

$$f(\mathbf{c}, T) = \min_{\mathbf{u}(0 \le t < \Delta t)} \min_{\mathbf{u}(\Delta t \le t \le T)} \left[\lambda \int_0^{\Delta t} \mathbf{u}^*(t)\mathbf{u}(t)\, dt + \mathbf{x}^*(T)\mathbf{P}\mathbf{x}(T) + \lambda \int_{\Delta t}^T \mathbf{u}^*(t)\mathbf{u}(t)\, dt \right] \tag{9–20}$$

Equation (9–20) can be approximated by

$$f(\mathbf{c}, T) = \min_{\mathbf{u}} \left[\lambda \int_0^{\Delta t} \mathbf{u}^*(t)\mathbf{u}(t)\, dt + f(\mathbf{c} + \Delta\mathbf{c}, T - \Delta t) \right]$$

The partial differential equation that must be satisfied by $f(\mathbf{c}, T)$ can be obtained as

$$\frac{\partial f}{\partial T} = \min_{\mathbf{u}} \left[\lambda \mathbf{u}^*(0)\mathbf{u}(0) + (\nabla h)^*[\mathbf{A}\mathbf{c} + \mathbf{B}\mathbf{u}(0)] \right] \tag{9–21}$$

where $(\nabla h)^* = (\nabla f)' = (\text{grad}\, f)'$

Note that if $\mathbf{c}$ is a real vector, then $\nabla h = \nabla f$. If $\mathbf{c}$ is a complex vector, then $\nabla h = \overline{\nabla f}$. Note also that we can write

$$\lambda \mathbf{u}^*(0)\mathbf{u}(0) + (\nabla h)^*[\mathbf{A}\mathbf{c} + \mathbf{B}\mathbf{u}(0)] = \frac{1}{\lambda}\left[\lambda \mathbf{u}(0) + \frac{\mathbf{B}^*}{2}\nabla h\right]^*\left[\lambda \mathbf{u}(0) + \frac{\mathbf{B}^*}{2}\nabla h\right] + (\nabla h)^*\mathbf{A}\mathbf{c} - \frac{(\nabla h)^*}{4\lambda}\mathbf{B}\mathbf{B}^*\nabla h \tag{9–22}$$

Let us denote the optimal control vector, the control vector $\mathbf{u}(0)$ that minimizes the left-hand side of Eq. (9–22), as $\mathbf{u}_0(0)$. Then, clearly the optimal control vector $\mathbf{u}_0(0)$ is given by

$$\mathbf{u}_0(0) = -\frac{1}{2\lambda}\mathbf{B}^*\nabla h \tag{9–23}$$

Substituting Eq. (9–23) into Eq. (9–21),

$$\frac{\partial f}{\partial T} = \lambda\left(-\frac{1}{2\lambda}\mathbf{B}^*\nabla h\right)^*\left(-\frac{1}{2\lambda}\mathbf{B}^*\nabla h\right) + (\nabla h)^*\left[\mathbf{A}\mathbf{c} + \mathbf{B}\left(-\frac{1}{2\lambda}\mathbf{B}^*\nabla h\right)\right]$$

which becomes

$$\frac{\partial f}{\partial T} = -\frac{1}{4\lambda}(\nabla h)^*\mathbf{B}\mathbf{B}^*\nabla h + (\nabla h)^*\mathbf{A}\mathbf{c}$$

Since the performance index is quadratic or Hermitian, a quadratic or Hermitian solution of the form

$$f(\mathbf{c}, T) = \mathbf{c}^*\mathbf{S}(T)\mathbf{c}$$

is assumed. (In Section 9–2 we have shown that $f(\mathbf{c}, T)$ is actually of a quadratic or Hermitian form.) Since $f(\mathbf{c}, T)$ is a real scalar, $f^* = f$, or

$$\mathbf{c}^*\mathbf{S}(T)\mathbf{c} = f(\mathbf{c}, T) = f^*(\mathbf{c}, T) = \mathbf{c}^*\mathbf{S}^*(T)\mathbf{c}$$

Thus we can take $\mathbf{S}(T)$ to be real symmetric or Hermitian. By use of the relationships

$$\frac{\partial f}{\partial T} = \mathbf{c}^*\frac{d\mathbf{S}}{dT}\mathbf{c} \qquad \text{and} \qquad (\nabla h)^* = (\nabla f)' = 2\mathbf{c}^*\mathbf{S}$$

we obtain

$$\mathbf{c}^*\frac{d\mathbf{S}}{dT}\mathbf{c} = -\frac{1}{4\lambda}(2\mathbf{c}^*\mathbf{S})\mathbf{B}\mathbf{B}^*(2\mathbf{S}\mathbf{c}) + 2\mathbf{c}^*\mathbf{S}\mathbf{A}\mathbf{c} = \mathbf{c}^*\left[-\frac{1}{\lambda}\mathbf{S}\mathbf{B}\mathbf{B}^*\mathbf{S} + \mathbf{S}\mathbf{A} + \mathbf{A}^*\mathbf{S}\right]\mathbf{c}$$

or, rearranging,

$$\mathbf{c}^*\left[\frac{d\mathbf{S}}{dT} - \mathbf{S}\mathbf{A} - \mathbf{A}^*\mathbf{S} + \frac{1}{\lambda}\mathbf{S}\mathbf{B}\mathbf{B}^*\mathbf{S}\right]\mathbf{c} = 0$$

This last equation is satisfied if $\mathbf{S}$ satisfies the following matrix Riccati equation:

$$\frac{d\mathbf{S}}{dT} - \mathbf{SA} - \mathbf{A}^*\mathbf{S} + \frac{1}{\lambda}\mathbf{SBB}^*\mathbf{S} = \mathbf{0} \tag{9–24}$$

For the solution of Eq. (9–24), let

$$\mathbf{S}(T) = \mathbf{Z}(T)\mathbf{Y}^{-1}(T)$$

where $\mathbf{Y}(T)$ and $\mathbf{Z}(T)$ satisfy the following equations

$$\frac{d\mathbf{Y}}{dT} = -\mathbf{AY} + \frac{1}{\lambda}\mathbf{BB}^*\mathbf{Z}$$

$$\frac{d\mathbf{Z}}{dT} = \mathbf{A}^*\mathbf{Z}$$

Since the solution of the differential equation

$$\frac{d\mathbf{x}}{dt} = \mathbf{Ax} + \mathbf{Bu}, \qquad \mathbf{x}(0) = \mathbf{c}$$

is given by

$$\mathbf{x}(T) = e^{\mathbf{A}T}\mathbf{c} + \int_0^T e^{\mathbf{A}(T-s)}\mathbf{Bu}(s)\,ds$$

the solution for $\mathbf{Z}(T)$ is obtained as

$$\mathbf{Z}(T) = (e^{\mathbf{A}T})^*\mathbf{Z}(0)$$

and the solution for $\mathbf{Y}(T)$ is

$$\mathbf{Y}(T) = e^{-\mathbf{A}T}\mathbf{Y}(0) + \int_0^T e^{-\mathbf{A}(T-s)}\frac{1}{\lambda}\mathbf{BB}^*\mathbf{Z}(s)\,ds$$

$$= \frac{1}{\lambda}e^{-\mathbf{A}T}\left[\lambda\mathbf{Y}(0) + \int_0^T e^{\mathbf{A}s}\mathbf{BB}^*(e^{\mathbf{A}s})^*\mathbf{Z}(0)\,ds\right]$$

The solution $\mathbf{S}(T)$ requires $\mathbf{Y}^{-1}(T)$. Hence inverting $\mathbf{Y}(T)$

$$\mathbf{Y}^{-1}(T) = \lambda\left[\lambda\mathbf{Y}(0) + \int_0^T e^{\mathbf{A}s}\mathbf{BB}^*(e^{\mathbf{A}s})^*\mathbf{Z}(0)\,ds\right]^{-1}e^{\mathbf{A}T}$$

The initial conditions are determined as

$$f(\mathbf{c}, 0) = \mathbf{c}^*\mathbf{Pc} = \mathbf{c}^*\mathbf{S}(0)\mathbf{c}$$

Hence,

$$\mathbf{S}(0) = \mathbf{P}$$

Also we have the relationship $\mathbf{S}(T) = \mathbf{Z}(T)\mathbf{Y}^{-1}(T)$ which yields

$$\mathbf{S}(0) = \mathbf{Z}(0)\,\mathbf{Y}^{-1}(0)$$

Let us choose $\mathbf{Z}(0) = \mathbf{P}$ and $\mathbf{Y}^{-1}(0) = \mathbf{I}$ so that $\mathbf{Y}^{-1}(T)$ exists. Then

$$\mathbf{S}(T) = \lambda(e^{\mathbf{A}T})^*\mathbf{P}\left[\lambda\mathbf{I} + \int_0^T e^{\mathbf{A}s}\mathbf{BB}^*(e^{\mathbf{A}s})^*\mathbf{P}\,ds\right]^{-1}e^{\mathbf{A}T}$$

It can be seen that $\mathbf{S}(T)$ is real symmetric or Hermitian. $f(\mathbf{c}, T)$ is obtained as

$$f(\mathbf{c}, T) = \mathbf{c}^*\mathbf{S}(T)\mathbf{c}$$

$$= \mathbf{c}^*\left\{\lambda(e^{\mathbf{A}T})^*\mathbf{P}\left[\lambda\mathbf{I} + \int_0^T e^{\mathbf{A}s}\mathbf{BB}^*(e^{\mathbf{A}s})^*\mathbf{P}\,ds\right]^{-1}e^{\mathbf{A}T}\right\}\mathbf{c} \tag{9–25}$$

The equation of the optimal control vector $\mathbf{u}_0(0)$ is given by Eq. (9–23), rewritten

$$\mathbf{u}_0(0) = -\frac{1}{2\lambda}\mathbf{B}^*\,\nabla h$$

Noting that

$$\nabla h = 2\mathbf{S}(T)\mathbf{c}$$

the optimal control vector $\mathbf{u}_0(0)$ becomes

$$\mathbf{u}_0(0) = -\frac{1}{2\lambda}\mathbf{B}^*2\mathbf{S}(T)\mathbf{c}$$

$$= -\mathbf{B}^*(e^{\mathbf{A}T})^*\mathbf{P}\left[\lambda\mathbf{I} + \int_0^T e^{\mathbf{A}s}\mathbf{BB}^*(e^{\mathbf{A}s})^*\mathbf{P}\,ds\right]^{-1}e^{\mathbf{A}T}\mathbf{c} \tag{9–26}$$

Letting $T - t$ be the duration of time remaining in the process and $\mathbf{x}(t)$ be the current state, the equation of the optimal control vector becomes

$$\mathbf{u}_0(t) = -\mathbf{B}^*(e^{\mathbf{A}(T-t)})^*\mathbf{P}\left[\lambda\mathbf{I} + \int_0^{T-t} e^{\mathbf{A}s}\mathbf{BB}^*(e^{\mathbf{A}s})^*\mathbf{P}\,ds\right]^{-1}e^{\mathbf{A}(T-t)}\mathbf{x}(t)$$

which is identical to Eq. (9–12). The minimized performance index can be given by

$$f(\mathbf{x}(t), T - t) = \mathbf{x}^*(t)\mathbf{S}(T - t)\mathbf{x}(t)$$

$$= \mathbf{x}^*(t)\left\{\lambda(e^{\mathbf{A}(T-t)})^*\mathbf{P}\left[\lambda\mathbf{I} + \int_0^{T-t} e^{\mathbf{A}s}\mathbf{BB}^*(e^{\mathbf{A}s})^*\mathbf{P}\,ds\right]^{-1}e^{\mathbf{A}(T-t)}\right\}\mathbf{x}(t)$$

which is the same as Eq. (9–15). The solution thus obtained by use of the functional equations technique of dynamic programming is, as a matter of course, the same as that obtained in Section 9–2 by the direct method.

9-4 ANALYTICAL DESIGN OF MINIMUM ERROR REGULATORS

In this section we consider the problem of selecting the control vector from an allowable set so as to minimize a function of the system error vector over the duration of the process. The *system error* is the deviation from the ideal path. The deviation from the ideal can be measured in many different ways. In this section the *system error* is defined as the difference between the desired state which is the origin of the state space and the instantaneous state during the control process. In what follows we shall consider the case where

we wish to minimize the generalized quadratic or Hermitian error, subject to the constraint that the energy for control is limited.

Assume that the differential equation of the plant is given by

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \qquad \mathbf{x}(0) = \mathbf{c} \tag{9-27}$$

where
$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{A} = n \times n$ matrix
$\mathbf{B} = n \times r$ matrix

In the present regulator problem the system has no input for $t \geqslant 0$. (The system is subjected to only the initial condition.) The desired state throughout the time interval $0 \leqslant t \leqslant T$, where T is the duration of time in the process, is the origin of the state space. Thus the system error vector is the negative of the state vector. The performance index is given by

$$J(\mathbf{c}, T) = \mathbf{x}^*(T)\mathbf{P}\mathbf{x}(T) + \int_0^T \mathbf{x}^*(t)\mathbf{Q}(t)\mathbf{x}(t)\,dt + \int_0^T \mathbf{u}^*(t)\mathbf{R}(t)\mathbf{u}(t)\,dt \tag{9-28}$$

where
$\mathbf{P}$ = positive definite (or positive semidefinite) real symmetric or Hermitian matrix such that $p_{ii} = p_{jj}$ if x_i and x_j are a complex conjugate pair
$\mathbf{Q}(t)$ = positive definite (or positive semidefinite) real symmetric or Hermitian matrix such that $q_{ii} = q_{jj}$ if x_i and x_j are a complex conjugate pair
$\mathbf{R}(t)$ = positive definite real symmetric or Hermitian matrix

The matrix $\mathbf{P}$ determines how the terminal errors are to be weighted. The matrix $\mathbf{Q}(t)$ determines how the instantaneous errors are to be weighted during the control process. The matrix $\mathbf{R}(t)$ represents the weight of control cost with respect to the minimizing error. In the following material, this problem will be solved by the method based on the functional equations technique of dynamic programming and by the direct method.

(1) **Solution of minimum error regulator problem by the functional equations technique of dynamic programming** *(9-6, 9-7)* Let $f(\mathbf{c}, T)$ be the minimized performance index. Then from Eq. (9-28) we obtain

$$f(\mathbf{c}, T) = \min_{\mathbf{u}} J(\mathbf{c}, T)$$

$$= \min_{\mathbf{u}} \left[\mathbf{x}^*(T)\mathbf{P}\mathbf{x}(T) + \int_0^T \mathbf{x}^*(t)\mathbf{Q}(t)\mathbf{x}(t)\,dt + \int_0^T \mathbf{u}^*(t)\mathbf{R}(t)\mathbf{u}(t)\,dt\right]$$

Splitting the interval of integration into two parts $0 \leqslant t < \Delta t$ and $\Delta t \leqslant t \leqslant T$, and applying the principle of optimality, we obtain

$$f(\mathbf{c}, T) = \min_{\mathbf{u}} \left[\int_0^{\Delta t} [\mathbf{x}^*(t)\mathbf{Q}(t)\mathbf{x}(t) + \mathbf{u}^*(t)\mathbf{R}(t)\mathbf{u}(t)]\,dt + f(\mathbf{c} + \Delta\mathbf{c}, T - \Delta t)\right]$$

from which we obtain, after expanding $f(\mathbf{c} + \Delta\mathbf{c}, T - \Delta t)$ into a Taylor series and dropping $f(\mathbf{c}, T)$ from both sides and dividing the remaining terms by Δt,

$$0 = \min_{\mathbf{u}} \left[\mathbf{x}^*\mathbf{Q}\mathbf{x} + \mathbf{u}^*\mathbf{R}\mathbf{u} + \sum_{i=1}^{n} \frac{\partial f}{\partial c_i}\frac{\Delta c_i}{\Delta t} - \frac{\partial f}{\partial T} + 0(\Delta t)\right]$$

In the limit as Δt approaches zero, this last equation becomes

$$\frac{\partial f}{\partial T} = \min_{\mathbf{u}} \left[\mathbf{c}^*\mathbf{Q}(0)\mathbf{c} + \mathbf{u}^*(0)\mathbf{R}(0)\mathbf{u}(0) + (\nabla h)^*[\mathbf{A}\mathbf{c} + \mathbf{B}\mathbf{u}(0)]\right] \tag{9-29}$$

where $(\nabla h)^* = (\nabla f)' = (\text{grad}\, f)'$

The right-hand side of Eq. (9-29) becomes minimum when

$$2\mathbf{R}(0)\,\mathbf{u}(0) + \mathbf{B}^*\,\nabla h = \mathbf{0}$$

The control vector $\mathbf{u}(0)$ satisfying this equation is the optimal control vector. Therefore, denoting the optimal control vector as $\mathbf{u}_0(0)$ we have

$$\mathbf{u}_0(0) = -\frac{\mathbf{R}^{-1}(0)}{2}\mathbf{B}^*\,\nabla h \tag{9-30}$$

Substituting Eq. (9-30) into Eq. (9-29)

$$\frac{\partial f}{\partial T} = \mathbf{c}^*\mathbf{Q}(0)\mathbf{c} + (\nabla h)^*\mathbf{A}\mathbf{c} - \frac{1}{4}(\nabla h)^*\mathbf{B}\mathbf{R}^{-1}(0)\mathbf{B}^*\,\nabla h \tag{9-31}$$

Assume the following quadratic or Hermitian form for the solution

$$f(\mathbf{c}, T) = \mathbf{c}^*\mathbf{S}(T)\mathbf{c} \tag{9-32}$$

where $\mathbf{S}(T)$ is real symmetric or Hermitian. From Eq. (9-31) and Eq. (9-32) noting that $\nabla h = 2\mathbf{S}(T)\mathbf{c}$

$$\mathbf{c}^*\frac{d\mathbf{S}}{dT}\mathbf{c} = \mathbf{c}^*[\mathbf{Q}(0) + \mathbf{S}\mathbf{A} + \mathbf{A}^*\mathbf{S} - \mathbf{S}\mathbf{B}\mathbf{R}^{-1}(0)\mathbf{B}^*\mathbf{S}]\mathbf{c} \tag{9-33}$$

For a solution of Eq. (9-33) find $\mathbf{S}$ which satisfies the following equation:

$$\frac{d\mathbf{S}}{dT} - \mathbf{S}\mathbf{A} - \mathbf{A}^*\mathbf{S} + \mathbf{S}\mathbf{B}\mathbf{R}^{-1}(0)\mathbf{B}^*\mathbf{S} = \mathbf{Q}(0) \tag{9-34}$$

This is a nonhomogeneous matrix Riccati equation. The solution of Eq. (9-34) may be obtained by letting

$$\mathbf{S}(T) = \mathbf{Z}(T)\mathbf{Y}^{-1}(T)$$

and solving the set of matrix differential equations:

$$\frac{d\mathbf{Y}(T_1)}{dT_1} = -\mathbf{A}\mathbf{Y}(T_1) + \mathbf{B}\mathbf{R}^{-1}(T - T_1)\mathbf{B}^*\mathbf{Z}(T_1)$$

$$\frac{d\mathbf{Z}(T_1)}{dT_1} = \mathbf{Q}(T - T_1)\mathbf{Y}(T_1) + \mathbf{A}^*\mathbf{Z}(T_1) \tag{9-35}$$

Consider the case where $\mathbf{Q}$ and $\mathbf{R}$ are constant matrices. Define

$$\mathbf{W}(T) = \begin{bmatrix} \mathbf{Y}(T) \\ \hline \mathbf{Z}(T) \end{bmatrix} \quad \text{and} \quad \mathbf{M} = \left[\begin{array}{c|c} -\mathbf{A} & \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^* \\ \hline \mathbf{Q} & \mathbf{A}^* \end{array}\right]$$

Then Eq. (9–35) becomes

$$\frac{d\mathbf{W}(T)}{dT} = \mathbf{M}\mathbf{W}(T) \tag{9–36}$$

The initial condition for this matrix differential equation is obtained as follows: Since

$$f(\mathbf{c}, 0) = \mathbf{c}^*\mathbf{S}(0)\mathbf{c} = \mathbf{c}^*\mathbf{P}\mathbf{c}$$

we obtain

$$\mathbf{S}(0) = \mathbf{P}$$

Noting that

$$\mathbf{S}(0) = \mathbf{Z}(0)\,\mathbf{Y}^{-1}(0)$$

we select $\mathbf{Z}(0) = \mathbf{P}$ and $\mathbf{Y}(0) = \mathbf{I}$ so that the inverse of $\mathbf{Y}(T)$ exists. Thus the initial condition of Eq. (9-36) is given by

$$\mathbf{W}(0) = \begin{bmatrix} \mathbf{I} \\ \hline \mathbf{P} \end{bmatrix}$$

The solution of Eq. (9–36) is then obtained as

$$\mathbf{W}(T) = e^{\mathbf{M}T}\mathbf{W}(0)$$

If we express $e^{\mathbf{M}T}$ in the following $2n \times 2n$ matrix

$$e^{\mathbf{M}T} = \begin{bmatrix} \Phi_{11}(T) & \Phi_{12}(T) \\ \hline \Phi_{21}(T) & \Phi_{22}(T) \end{bmatrix}, \qquad \Phi_{ij}(T) = n \times n \text{ matrix}$$

$$(i, j = 1, 2)$$

then,

$$\begin{bmatrix} \mathbf{Y}(T) \\ \hline \mathbf{Z}(T) \end{bmatrix} = \begin{bmatrix} \Phi_{11}(T) & \Phi_{12}(T) \\ \hline \Phi_{21}(T) & \Phi_{22}(T) \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ \hline \mathbf{P} \end{bmatrix} \tag{9–37}$$

or

$$\mathbf{Y}(T) = \Phi_{11}(T) + \Phi_{12}(T)\mathbf{P}$$
$$\mathbf{Z}(T) = \Phi_{21}(T) + \Phi_{22}(T)\mathbf{P}$$

Hence

$$\begin{aligned}\mathbf{S}(T) &= \mathbf{Z}(T)\mathbf{Y}^{-1}(T) \\ &= [\Phi_{21}(T) + \Phi_{22}(T)\mathbf{P}]\,[\Phi_{11}(T) + \Phi_{12}(T)\mathbf{P}]^{-1}\end{aligned}$$

and

$$\begin{aligned}\mathbf{u}_0(0) &= -\mathbf{R}^{-1}\mathbf{B}^*\mathbf{S}(T)\mathbf{c} = -\mathbf{R}^{-1}\mathbf{B}^*\mathbf{Z}(T)\mathbf{Y}^{-1}(T)\mathbf{c} \\ &= -\mathbf{R}^{-1}\mathbf{B}^*[\Phi_{21}(T) + \Phi_{22}(T)\mathbf{P}]\,[\Phi_{11}(T) + \Phi_{12}(T)\mathbf{P}]^{-1}\mathbf{c}\end{aligned}$$

$f(\mathbf{c}, T)$ is obtained as

$$\begin{aligned}f(\mathbf{c}, T) &= \mathbf{c}^*\mathbf{S}(T)\mathbf{c} = \mathbf{c}^*\mathbf{Z}(T)\mathbf{Y}^{-1}(T)\mathbf{c} \\ &= \mathbf{c}^*[\Phi_{21}(T) + \Phi_{22}(T)\mathbf{P}]\,[\Phi_{11}(T) + \Phi_{12}(T)\mathbf{P}]^{-1}\mathbf{c}\end{aligned} \tag{9–38}$$

Let $T - t$ be the duration of time remaining and $\mathbf{x}(t)$ be the current state. Then the optimal control vector [noting that $\Phi_{11}(T)$, $\Phi_{12}(T)$, $\Phi_{21}(T)$, and $\Phi_{22}(T)$ are functions of T] becomes

$$\mathbf{u}_0(t) = -\mathbf{R}^{-1}\mathbf{B}^*\mathbf{S}(T - t)\mathbf{x}(t) = -\mathbf{F}(T - t)\mathbf{x}(t) \tag{9–39}$$

where

$$\mathbf{F}(T-t) = \mathbf{R}^{-1}\mathbf{B}^*[\Phi_{21}(T-t) + \Phi_{22}(T-t)\mathbf{P}][\Phi_{11}(T-t) + \Phi_{12}(T-t)\mathbf{P}]^{-1}$$
$$0 \leqslant t \leqslant T$$

and

$$\begin{aligned}f(\mathbf{x}(t), T - t) &= \mathbf{x}^*(t)\mathbf{S}(T - t)\mathbf{x}(t) = \mathbf{x}^*(t)\mathbf{Z}(T - t)\mathbf{Y}^{-1}(T - t)\mathbf{x}(t) \\ &= \mathbf{x}^*(t)[\Phi_{21}(T - t) + \Phi_{22}(T - t)\mathbf{P}]\,[\Phi_{11}(T - t) + \Phi_{12}(T - t)\mathbf{P}]^{-1}\mathbf{x}(t)\end{aligned}$$

The block diagram of the optimal control system is shown in Fig. 9–9. The system designed (if $|\mathbf{P}| > 0$) is asymptotically stable during the process of control, $0 \leqslant t \leqslant T$, since $f(\mathbf{c}, T)$ is a Liapunov function of the system. As for stability of the system for $t > T$, the analysis made in Section 9–2 also holds true in this case.

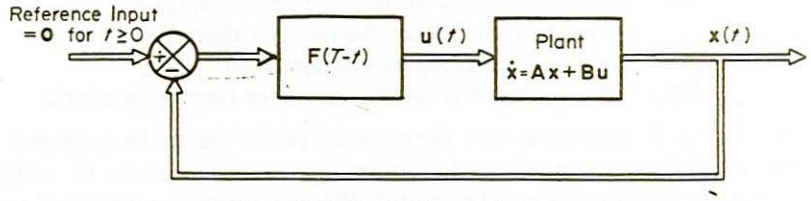


Figure 9-9 Optimal control system.

(2) **Solution of minimum error regulator problem by the direct method** (*9–1*) The solution to Eq. (9–27) can be obtained as

$$\mathbf{x}(s) = e^{\mathbf{A}s}\left[\mathbf{c} + \int_0^s e^{-\mathbf{A}t}\mathbf{B}\mathbf{u}(t)\,dt\right]$$

Denote by $J_0(\mathbf{c}, T)$ and $\mathbf{x}_0(s)$ the J and $\mathbf{x}$ corresponding to a special $\mathbf{u}_0$. The control vector $\mathbf{u}_0$ will be specified later. Since

$$\mathbf{x}_0(s) = e^{\mathbf{A}s}\left[\mathbf{c} + \int_0^s e^{-\mathbf{A}t}\mathbf{B}\mathbf{u}_0(t)\,dt\right]$$

we have

$$\mathbf{x}(s) - \mathbf{x}_0(s) = \int_0^s e^{\mathbf{A}(s-t)}\mathbf{B}[\mathbf{u}(t) - \mathbf{u}_0(t)]\,dt$$

Define $\mathbf{y}(t) = \mathbf{x}(t) - \mathbf{x}_0(t)$. Then

$$J(\mathbf{c}, T) = J_0(\mathbf{c}, T) + \mathbf{y}^*(T)\mathbf{P}\mathbf{y}(T) + \int_0^T \mathbf{y}^*(t)\mathbf{Q}(t)\mathbf{y}(t)\,dt + \int_0^T [\mathbf{u}(t) - \mathbf{u}_0(t)]^*\mathbf{R}(t)[\mathbf{u}(t) - \mathbf{u}_0(t)]\,dt + \mathbf{x}_0^*(T)\mathbf{P}\mathbf{y}(T) + \mathbf{y}^*(T)\mathbf{P}\mathbf{x}_0(T) + \int_0^T \mathbf{x}_0^*(t)\mathbf{Q}(t)\mathbf{y}(t)\,dt + \int_0^T \mathbf{y}^*(t)\mathbf{Q}(t)\mathbf{x}_0(t)dt + \int_0^T \mathbf{u}_0^*(t)\mathbf{R}(t)[\mathbf{u}(t) - \mathbf{u}_0(t)]\,dt + \int_0^T [\mathbf{u}(t) - \mathbf{u}_0(t)]^*\mathbf{R}(t)\mathbf{u}_0(t)dt \tag{9-40}$$

Since

$$\mathbf{x}_0^*(T)\mathbf{P}\mathbf{y}(T) = \mathbf{x}_0^*(T)\mathbf{P}e^{\mathbf{A}T}\int_0^T e^{-\mathbf{A}t}\mathbf{B}[\mathbf{u}(t) - \mathbf{u}_0(t)]\,dt$$

and

$$\int_0^T \mathbf{x}_0^*(s)\mathbf{Q}(s)\mathbf{y}(s)\,ds = \int_0^T \mathbf{x}_0^*(s)\mathbf{Q}(s)e^{\mathbf{A}s}\left[\int_0^s e^{-\mathbf{A}t}\mathbf{B}[\mathbf{u}(t) - \mathbf{u}_0(t)]\,dt\right]ds = \int_0^T \left(\int_t^T \mathbf{x}_0^*(s)\mathbf{Q}(s)e^{\mathbf{A}s}\,ds\right)e^{-\mathbf{A}t}\mathbf{B}[\mathbf{u}(t) - \mathbf{u}_0(t)]\,dt$$

where we have used the relationship

$$\int_0^T f(s)\left(\int_0^s g(t)dt\right)ds = \int_0^T\left(\int_t^T f(s)ds\right)g(t)dt$$

the sum of the last six terms of the right-hand side of Eq. (9–40) can be written as

$$\int_0^T \mathbf{U}^*(t)[\mathbf{u}(t) - \mathbf{u}_0(t)]\,dt + \int_0^T [\mathbf{u}(t) - \mathbf{u}_0(t)]^*\mathbf{U}(t)dt$$

where

$$\mathbf{U}(t) = \mathbf{B}^*(e^{-\mathbf{A}t})^*(e^{\mathbf{A}T})^*\mathbf{P}\mathbf{x}_0(T) + \mathbf{B}^*(e^{-\mathbf{A}t})^*\int_t^T (e^{\mathbf{A}s})^*\mathbf{Q}(s)\mathbf{x}_0(s)\,ds + \mathbf{R}(t)\mathbf{u}_0(t)$$

We shall determine $\mathbf{u}_0(t)$ by solving the equation $\mathbf{U}(t) = \mathbf{0}$. With this $\mathbf{u}_0(t)$ the sum of the last six terms of the right-hand side of Eq. (9–40) is zero. Noting that $\mathbf{P}$ and $\mathbf{Q}(t)$ are positive definite or positive semidefinite and $\mathbf{R}(t)$ is positive definite, we see that $\mathbf{u}_0(t)$ gives the minimum of $J(\mathbf{c}, T)$. Hence $\mathbf{u}_0(t)$ is the optimal control vector. It is given by

$$\mathbf{u}_0(t) = -\mathbf{R}^{-1}(t)\mathbf{B}^*(e^{-\mathbf{A}t})^*\left[(e^{\mathbf{A}T})^*\mathbf{P}\mathbf{x}_0(T) + \int_t^T (e^{\mathbf{A}s})^*\mathbf{Q}(s)\mathbf{x}_0(s)\,ds\right] \tag{9-41}$$

Since

$$\mathbf{x}_0(s) = e^{\mathbf{A}(s-T)}\mathbf{x}_0(T) - \int_s^T e^{\mathbf{A}(s-t)}\mathbf{B}\mathbf{u}_0(t)\,dt \tag{9-42}$$

we have

$$\mathbf{x}_0(s) = \left[e^{\mathbf{A}(s-T)} + \int_s^T e^{\mathbf{A}(s-t)}\mathbf{B}\mathbf{R}^{-1}(t)\mathbf{B}^*(e^{-\mathbf{A}t})^*\,dt\,(e^{\mathbf{A}T})^*\mathbf{P}\right]\mathbf{x}_0(T) + \int_s^T e^{\mathbf{A}(s-t)}\mathbf{B}\mathbf{R}^{-1}(t)\mathbf{B}^*(e^{-\mathbf{A}t})^*\left(\int_t^T (e^{\mathbf{A}\sigma})^*\mathbf{Q}(\sigma)\mathbf{x}_0(\sigma)\,d\sigma\right)dt$$

Therefore, we have an integral equation

$$\mathbf{x}_0(s) = \mathbf{C}(s)\mathbf{x}_0(T) + \int_s^T \mathbf{D}(s, \sigma)\mathbf{x}_0(\sigma)\,d\sigma \tag{9-43}$$

where

$$\mathbf{C}(s) = e^{\mathbf{A}(s-T)} + \int_s^T e^{\mathbf{A}(s-t)}\mathbf{B}\mathbf{R}^{-1}(t)\mathbf{B}^*(e^{-\mathbf{A}t})^*\,dt\,(e^{\mathbf{A}T})^*\mathbf{P}$$

$$\mathbf{D}(s, \sigma) = e^{\mathbf{A}s}\int_s^\sigma e^{-\mathbf{A}t}\mathbf{B}\mathbf{R}^{-1}(t)\mathbf{B}^*\,(e^{-\mathbf{A}t})^*\,dt\,(e^{\mathbf{A}\sigma})^*\,\mathbf{Q}(\sigma)$$

where we have used the relationship

$$\int_s^T f(t)\left(\int_t^T g(\sigma)d\sigma\right)dt = \int_s^T\left(\int_s^\sigma f(t)dt\right)g(\sigma)d\sigma$$

Let

$$\mathbf{x}_0(s) = \mathbf{G}(s)\mathbf{x}_0(T) \tag{9-44}$$

where $\mathbf{G}(s)$ is an $n \times n$ matrix. From Eq. (9–43) and Eq. (9–44),

$$\mathbf{G}(s)\mathbf{x}_0(T) = \mathbf{C}(s)\mathbf{x}_0(T) + \int_s^T \mathbf{D}(s, \sigma)\mathbf{G}(\sigma)\mathbf{x}_0(T)\,d\sigma$$

Now consider the integral equation

$$\mathbf{G}(s) = \mathbf{C}(s) + \int_s^T \mathbf{D}(s, \sigma)\mathbf{G}(\sigma)\,d\sigma \tag{9-45}$$

where $\mathbf{G}(T) = \mathbf{I}$ and $\mathbf{C}(T) = \mathbf{I}$. There is a unique solution of Eq. (9–45). Let us put

$$\mathbf{u}_0(t) = \mathbf{L}(t)\mathbf{x}_0(T) \tag{9-46}$$

where

$$\mathbf{L}(t) = -\mathbf{R}^{-1}(t)\mathbf{B}^*(e^{-\mathbf{A}t})^*\left[(e^{\mathbf{A}T})^*\mathbf{P} + \int_t^T (e^{\mathbf{A}s})^*\mathbf{Q}(s)\mathbf{G}(s)\,ds\right]$$

Consider next the equation

$$\mathbf{x}_0(T) = e^{\mathbf{A}T}\left[\mathbf{c} + \int_0^T e^{-\mathbf{A}t}\mathbf{B}\mathbf{u}_0(t)\,dt\right] = e^{\mathbf{A}T}\left[\mathbf{c} + \int_0^T e^{-\mathbf{A}t}\mathbf{B}\mathbf{L}(t)\,dt\,\mathbf{x}_0(T)\right] \tag{9-47}$$

Solving Eq. (9–47) we get

$$\mathbf{x}_0(T) = \left[\mathbf{I} - e^{\mathbf{A}T}\int_0^T e^{-\mathbf{A}t}\mathbf{B}\mathbf{L}(t)\,dt\right]^{-1} e^{\mathbf{A}T}\mathbf{c} \tag{9–48}$$

Equation (9–46) and Eq. (9–48) define the optimal control vector $\mathbf{u}_0(t)$.

In the present approach, we need to solve the integral equation (9–45) to determine $\mathbf{u}_0(t)$. The problem of solving this integral equation, however, can be converted to that of solving the matrix Riccati equation (9–34) or its equivalent matrix differential equations (9–35). We shall show this in what follows. Since Eq. (9–45) can be written as

$$\mathbf{G}(s) = e^{\mathbf{A}(s-T)} + \int_s^T e^{\mathbf{A}(s-t)}\mathbf{B}\mathbf{R}^{-1}(t)\mathbf{B}^*(e^{-\mathbf{A}t})^*\,dt\;(e^{\mathbf{A}T})^*\mathbf{P}$$
$$+ \int_s^T e^{\mathbf{A}(s-t)}\mathbf{B}\mathbf{R}^{-1}(t)\mathbf{B}^*(e^{-\mathbf{A}t})^*\left[\int_t^T (e^{\mathbf{A}\sigma})^*\mathbf{Q}(\sigma)\mathbf{G}(\sigma)\,d\sigma\right]dt$$

we have

$$\mathbf{G}(s) = e^{\mathbf{A}(s-T)} + \int_s^T e^{\mathbf{A}(s-t)}\mathbf{B}\mathbf{R}^{-1}(t)\mathbf{B}^*\mathbf{K}(t)\,dt$$

where

$$\mathbf{K}(s) = (e^{-\mathbf{A}s})^*\left[(e^{\mathbf{A}T})^*\mathbf{P} + \int_s^T (e^{\mathbf{A}t})^*\mathbf{Q}(t)\mathbf{G}(t)\,dt\right]$$

Differentiating $\mathbf{G}(s)$ and $\mathbf{K}(s)$ with respect to s, we obtain

$$\frac{d\mathbf{G}(s)}{ds} = \mathbf{A}\mathbf{G}(s) - \mathbf{B}\mathbf{R}^{-1}(s)\mathbf{B}^*\mathbf{K}(s), \qquad \mathbf{G}(T) = \mathbf{I}$$

and

$$\frac{d\mathbf{K}(s)}{ds} = -\mathbf{Q}(s)\mathbf{G}(s) - \mathbf{A}^*\mathbf{K}(s), \qquad \mathbf{K}(T) = \mathbf{P}$$

Note that $T - s$ is the remaining time in the process. Define

$$\begin{aligned} \mathbf{Y}(T-s) &= \mathbf{G}(s), \qquad \mathbf{Y}(0) = \mathbf{G}(T) = \mathbf{I} \\ \mathbf{Z}(T-s) &= \mathbf{K}(s), \qquad \mathbf{Z}(0) = \mathbf{K}(T) = \mathbf{P} \end{aligned} \tag{9–49}$$

then,

$$\frac{d\mathbf{G}(s)}{ds} = -\frac{d\mathbf{Y}(T-s)}{d(T-s)}, \qquad \frac{d\mathbf{K}(s)}{ds} = -\frac{d\mathbf{Z}(T-s)}{d(T-s)}$$

and

$$\frac{d\mathbf{Y}(T-s)}{d(T-s)} = -\mathbf{A}\mathbf{Y}(T-s) + \mathbf{B}\mathbf{R}^{-1}(s)\mathbf{B}^*\mathbf{Z}(T-s)$$
$$\frac{d\mathbf{Z}(T-s)}{d(T-s)} = \mathbf{Q}(s)\mathbf{Y}(T-s) + \mathbf{A}^*\mathbf{Z}(T-s)$$

By letting $T - s = T_1$, we obtain

$$\begin{aligned} \frac{d\mathbf{Y}(T_1)}{dT_1} &= -\mathbf{A}\mathbf{Y}(T_1) + \mathbf{B}\mathbf{R}^{-1}(T-T_1)\mathbf{B}^*\mathbf{Z}(T_1) \\ \frac{d\mathbf{Z}(T_1)}{dT_1} &= \mathbf{Q}(T-T_1)\mathbf{Y}(T_1) + \mathbf{A}^*\mathbf{Z}(T_1) \end{aligned} \tag{9–50}$$

which are the same as Eq. (9–35). We have thus shown that the problem of solving Eq. (9–45) can be converted to that of solving Eq. (9–35).

The value of $f(\mathbf{c}, T)$, the minimum value of $J(\mathbf{c}, T)$, can be obtained [without assuming the quadratic or Hermitian form given by Eq. (9–32)] as follows: By using the relationships

$$\mathbf{L}(t) = -\mathbf{R}^{-1}(t)\mathbf{B}^*\mathbf{K}(t)$$

and

$$\mathbf{K}^*(s)\mathbf{B}\mathbf{R}^{-1}(s)\mathbf{B}^* = \mathbf{G}^*(s)\mathbf{A}^* - \frac{d\mathbf{G}^*(s)}{ds}$$

we have

$$\begin{aligned} f(\mathbf{c}, T) &= J_0(\mathbf{c}, T) \\ &= \mathbf{x}_0^*(T)\mathbf{P}\mathbf{x}_0(T) + \int_0^T \mathbf{x}_0^*(s)\mathbf{Q}(s)\mathbf{x}_0(s)\,ds + \int_0^T \mathbf{u}_0^*(s)\mathbf{R}(s)\mathbf{u}_0(s)\,ds \\ &= \mathbf{x}_0^*(T)\left[\mathbf{P} + \int_0^T \mathbf{G}^*(s)\mathbf{Q}(s)\mathbf{G}(s)\,ds + \int_0^T \mathbf{L}^*(s)\mathbf{R}(s)\mathbf{L}(s)\,ds\right]\mathbf{x}_0(T) \\ &= \mathbf{x}_0^*(T)\left[\mathbf{P} - \int_0^T \mathbf{G}^*(s)\left\{\frac{d\mathbf{K}(s)}{ds} + \mathbf{A}^*\mathbf{K}(s)\right\}ds\right. \\ &\qquad \left. + \int_0^T \mathbf{K}^*(s)\mathbf{B}\mathbf{R}^{-1}(s)\mathbf{B}^*\mathbf{K}(s)\,ds\right]\mathbf{x}_0(T) \\ &= \mathbf{x}_0^*(T)\left[\mathbf{P} - \int_0^T \mathbf{G}^*(s)\frac{d\mathbf{K}(s)}{ds}\,ds - \int_0^T \frac{d\mathbf{G}^*(s)}{ds}\mathbf{K}(s)\,ds\right]\mathbf{x}_0(T) \\ &= \mathbf{x}_0^*(T)\left[\mathbf{P} - \int_0^T \frac{d}{ds}\{\mathbf{G}^*(s)\mathbf{K}(s)\}\,ds\right]\mathbf{x}_0(T) \\ &= \mathbf{x}_0^*(T)[\mathbf{G}^*(0)\mathbf{K}(0)]\mathbf{x}_0(T) \end{aligned}$$

Since

$$\mathbf{x}_0(s) = \mathbf{G}(s)\mathbf{x}_0(T)$$

we obtain

$$\mathbf{x}_0(0) = \mathbf{G}(0)\mathbf{x}_0(T) \qquad \text{and} \qquad \mathbf{x}_0^*(0) = \mathbf{x}_0^*(T)\mathbf{G}^*(0)$$

Hence,

$$\widehat{f(\mathbf{c}, T)} = J_0(\mathbf{c}, T) = \mathbf{x}_0^*(0)\mathbf{K}(0)\mathbf{G}^{-1}(0)\mathbf{x}_0(0) = \mathbf{c}^*\mathbf{K}(0)\mathbf{G}^{-1}(0)\mathbf{c}$$

Thus, $f(\mathbf{c}, T)$ becomes a quadratic form or a Hermitian form. In terms of $\mathbf{Y}(T)$ and $\mathbf{Z}(T)$ given by Eq. (9–49) and Eq. (9–50) we obtain

$$f(\mathbf{c}, T) = J_0(\mathbf{c}, T) = \mathbf{c}^*\mathbf{Z}(T)\mathbf{Y}^{-1}(T)\mathbf{c}$$

Example 9-3 Given the control system shown in Fig. 9–10 and the performance index

$$J(c, T) = x'(T)x(T) + \int_0^T x'(t)x(t)\,dt + \lambda \int_0^T u'(t)u(t)\,dt$$

find the optimal control function $u_0(t)$ and the minimized performance index $f(c, T) = \min_u J(c, T)$.

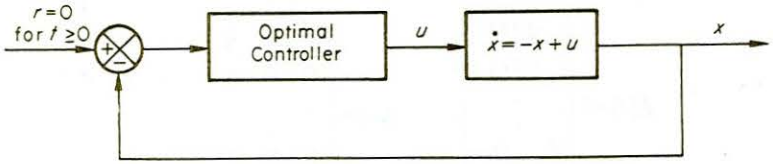


Figure 9-10 Control system.

The differential equation of the plant is

$$\dot{x} = Ax + Bu, \qquad x(0) = c$$

where x, u, A, and B are scalars. Clearly, $A = -1$ and $B = 1$. From the preceding analysis it can be shown that the optimal control function $u_0(t)$ and the minimized performance index $f(c, T)$ are representable as

$$u_0(t) = -\frac{1}{\lambda} B'S(T - t)x(t) \qquad \text{and} \qquad f(c, T) = c'S(T)c$$

where $$S(T) = [\phi_{21}(T) + \phi_{22}(T)]\,[\phi_{11}(T) + \phi_{12}(T)]^{-1}$$

and $\phi_{11}(T)$, $\phi_{12}(T)$, $\phi_{21}(T)$, and $\phi_{22}(T)$ are defined by

$$e^{\mathbf{M}T} = \begin{bmatrix} \phi_{11}(T) & \phi_{12}(T) \\ \hline \phi_{21}(T) & \phi_{22}(T) \end{bmatrix}$$

where

$$\mathbf{M} = \begin{bmatrix} -A & \frac{1}{\lambda} BB' \\ \hline 1 & A' \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{\lambda} \\ 1 & -1 \end{bmatrix}$$

The eigenvalues of $\mathbf{M}$ are found from

$$|\mathbf{M} - \xi\mathbf{I}| = \xi^2 - 1 - \frac{1}{\lambda} = 0$$

or

$$\xi_1 = \sqrt{\frac{\lambda + 1}{\lambda}}, \qquad \xi_2 = -\sqrt{\frac{\lambda + 1}{\lambda}}$$

Let us use Sylvester's interpolation formula to obtain $e^{\mathbf{M}T}$.

$$\begin{vmatrix} 1 & \xi_1 & e^{\xi_1 T} \\ 1 & \xi_2 & e^{\xi_2 T} \\ \mathbf{I} & \mathbf{M} & e^{\mathbf{M}T} \end{vmatrix} = 0$$

Noting that $\xi_2 = -\xi_1$ and expanding this determinant around the last column

$$e^{\mathbf{M}T} = \frac{-1}{2\xi_1} \begin{bmatrix} (1 - \xi_1)e^{-\xi_1 T} - (1 + \xi_1)e^{\xi_1 T} & \frac{1}{\lambda}(e^{-\xi_1 T} - e^{\xi_1 T}) \\ e^{-\xi_1 T} - e^{\xi_1 T} & (1 - \xi_1)e^{\xi_1 T} - (1 + \xi_1)e^{-\xi_1 T} \end{bmatrix}$$

Then,

$$\phi_{21}(T) + \phi_{22}(T) = \tfrac{1}{2}(e^{\xi_1 T} + e^{-\xi_1 T})$$

and

$$\phi_{11}(T) + \phi_{12}(T) = \frac{1}{2\xi_1}\left[\left(1 + \xi_1 + \frac{1}{\lambda}\right)e^{\xi_1 T} - \left(1 - \xi_1 + \frac{1}{\lambda}\right)e^{-\xi_1 T}\right]$$

or

$$[\phi_{11}(T) + \phi_{12}(T)]^{-1} = \frac{2\xi_1}{[1 + \xi_1 + (1/\lambda)]e^{\xi_1 T} - [1 - \xi_1 + (1/\lambda)]e^{-\xi_1 T}}$$

Thus

$$S(T) = [\phi_{21}(T) + \phi_{22}(T)]\,[\phi_{11}(T) + \phi_{12}(T)]^{-1}$$

$$= \frac{\xi_1(e^{\xi_1 T} + e^{-\xi_1 T})}{[1 + \xi_1 + (1/\lambda)]e^{\xi_1 T} - [1 - \xi_1 + (1/\lambda)]e^{-\xi_1 T}}$$

Substituting $\xi_1 = \sqrt{(\lambda + 1)/\lambda}$ into the preceding equation

$$S(T) = \frac{1 + e^{-2\sqrt{(\lambda+1)/\lambda}\,T}}{1 + \sqrt{(\lambda + 1)/\lambda} + [1 - \sqrt{(\lambda + 1)/\lambda}]e^{-2\sqrt{(\lambda+1)/\lambda}\,T}}$$

Hence

$$u_0(t) = \frac{-(1 + e^{-2\sqrt{(\lambda+1)/\lambda}\,(T-t)})x(t)}{\lambda + \sqrt{\lambda(\lambda + 1)} + [\lambda - \sqrt{\lambda(\lambda + 1)}]e^{-2\sqrt{(\lambda+1)/\lambda}\,(T-t)}}$$

and

$$f(c, T) = \frac{1 + e^{-2\sqrt{(\lambda+1)/\lambda}\,T}}{1 + \sqrt{(\lambda + 1)/\lambda} + [1 - \sqrt{(\lambda + 1)/\lambda}]e^{-2\sqrt{(\lambda+1)/\lambda}\,T}}\,c^2$$

Figure 9–11 shows a block diagram of the optimal control system designed.

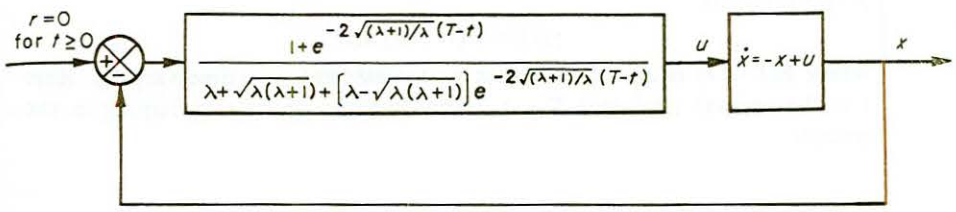


Figure 9-11 Block diagram for optimal control of plant defined by $\dot{x} = -x + u$.

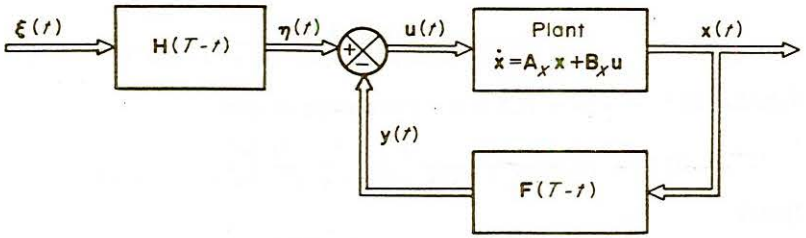
9-5 MINIMUM ERROR CONTROLLERS WITH POLYNOMIAL INPUTS (9-5, 9-6, 9-8)

The problems considered in Sections 9–2 to 9–4 are formulated as regulator problems; namely, the system has no input and is subjected to only the initial condition. In this section we shall consider systems with nonzero inputs which may be approximated by polynomials in t. In synthesizing a minimum error controller it is assumed that the polynomial input is known as a function of time for a given time interval $0 \leqslant t \leqslant T$. Note that if a system is designed so as to perform optimally with respect to some class of signals such as polynomials in t of degree m, then the same system will perform nearly optimally in response to any other input signal which can be approximated suitably by an mth-degree polynomial in t over some interval of time.

Consider the control system shown in Fig. 9–12. The state space equation of the plant is given by

$$\dot{\mathbf{x}} = \mathbf{A}_x\mathbf{x} + \mathbf{B}_x\mathbf{u}, \qquad \mathbf{x}(0) = \mathbf{c}$$

where
$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{A}_x = n \times n$ constant matrix
$\mathbf{B}_x = n \times r$ constant matrix



$$\eta(t) = \mathbf{H}(T-t)\,\xi(t), \qquad \mathbf{y}(t) = \mathbf{F}(T-t)\,\mathbf{x}(t)$$

Figure 9-12 Control system.

In the block diagram of Fig. 9–12 the feedback vector $\mathbf{y}(t)$ is assumed to be a linear function of the state vector $\mathbf{x}(t)$. The relationship between $\mathbf{x}(t)$ and $\mathbf{y}(t)$ is given by

$$\mathbf{y}(t) = \mathbf{F}(T-t)\,\mathbf{x}(t)$$

where $\mathbf{F}(T-t)$ is an $r \times n$ matrix whose elements are time-varying. Here t is the present time and $T-t$ is the duration of time remaining in the process.

Let $\{\bar{\xi}(t)\}$ be a class of system inputs for which the system is designed to be optimal. We assume that every member of the class $\{\bar{\xi}(t)\}$ satisfies the mth-order differential equation. We define an element $\bar{\xi}(t)$ of this class as

$$\bar{\xi}(t) = \xi_1(t) \qquad \text{for } t \geqslant 0$$
$$\bar{\xi}(t) = 0 \qquad \text{for } t < 0$$

where $\xi_1(t)$ is the first element of the input state vector $\boldsymbol{\xi}(t)$ which is defined by the following state space equation:

$$\dot{\boldsymbol{\xi}} = \mathbf{A}_\xi\boldsymbol{\xi}$$

where

$$\boldsymbol{\xi}(t) = \begin{bmatrix} \xi_1(t) \\ \xi_2(t) \\ \vdots \\ \xi_{m-1}(t) \\ \xi_m(t) \end{bmatrix}, \qquad \mathbf{A}_\xi = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

The first element $\xi_1(t)$ of the input state vector $\boldsymbol{\xi}(t)$ is the polynomial approximating the actual input signal. The matrix $\mathbf{A}_\xi$ is an $m \times m$ matrix. It is noted that, as far as the state variables of the input are concerned, these are strictly fictitious quantities. They have been introduced in order to permit optimizing system performance for a particular class of input signals. In the block diagram the vector $\boldsymbol{\eta}(t)$ is defined by the following relationship:

$$\boldsymbol{\eta}(t) = \mathbf{H}(T-t)\,\boldsymbol{\xi}(t)$$

where $\mathbf{H}(T-t)$ is an $r \times m$ matrix whose elements are time-varying.

The problem considered in the following is to determine, for a given polynomial input, the optimal controller or the elements of the matrices $\mathbf{F}(T-t)$ and $\mathbf{H}(T-t)$ such that the given system performance index is minimized. Such a problem may be solved by defining a new state vector $\mathbf{z}(t)$ consisting of $\mathbf{x}(t)$ and $\boldsymbol{\xi}(t)$ such that

$$\mathbf{z}(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \\ \hline \xi_1(t) \\ \vdots \\ \xi_m(t) \end{bmatrix} = \begin{bmatrix} \mathbf{x}(t) \\ \hline \boldsymbol{\xi}(t) \end{bmatrix} \tag{9-51}$$

Clearly, $\mathbf{z}(t)$ given by Eq. (9–51) is an $n + m$ vector. By defining

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{A}_x & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{A}_\xi \end{array}\right] \qquad \mathbf{B} = \left[\begin{array}{c} \mathbf{B}_x \\ \hline \mathbf{0} \end{array}\right] \tag{9–52}$$

the state space equation of the system can be written as

$$\dot{\mathbf{z}} = \mathbf{Az} + \mathbf{Bu}$$

Let us consider the quadratic or Hermitian performance index defined by

$$J(\mathbf{c}, T) = \mathbf{z}^*(T)\mathbf{P}\mathbf{z}(T) + \lambda \int_0^T \mathbf{u}^*(t)\mathbf{u}(t)dt$$

where $\mathbf{P}$ is an $(m + n) \times (m + n)$ positive definite (or positive semi-definite) real symmetric or Hermitian matrix characterizing the weight of errors specified for performance measurement, and λ is a positive constant which indicates the weight of control cost with respect to the minimizing errors. [Note that if x_i and x_j are a complex conjugate pair, then $p_{ii} = p_{jj}$ where $\mathbf{P} = (p_{ij})$.] Generally, $\mathbf{z}^*(T)\,\mathbf{P}\mathbf{z}(T)$ may be given by

$$\mathbf{z}^*(T)\mathbf{P}\mathbf{z}(T) = \sum_{i=1}^{m} w_i[\xi_i(T) - x_i(T)]^2$$

where the w_i are weighting factors, $\xi_1 - x_1$ is the error, $\xi_2 - x_2$ is the error rate, and so on. For example, if both $\mathbf{x}$ and $\boldsymbol{\xi}$ are two-dimensional vectors and if it is desired to minimize $[\xi_1(T) - x_1(T)]^2 + [\xi_2(T) - x_2(T)]^2$, then the matrix $\mathbf{P}$ is given by

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}$$

In fact,

$$\begin{aligned} &\mathbf{z}^*(T)\mathbf{P}\mathbf{z}(T) \\ &= [x_1(T)\ \ x_2(T)\ \ \xi_1(T)\ \ \xi_2(T)] \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(T) \\ x_2(T) \\ \xi_1(T) \\ \xi_2(T) \end{bmatrix} \\ &= [\xi_1(T) - x_1(T)]^2 + [\xi_2(T) - x_2(T)]^2 \end{aligned}$$

It is possible to assign different weights to the individual errors by appropriately selecting the elements of $\mathbf{P}$.

The derivation of the optimal control vector for this problem is identical with that of the problem discussed in Sections 9–2 and 9–3, where no input signal was considered. Therefore, the derivation of the optimal control vector to the present problem will not be repeated. From Eq. (9–12) the equation of the optimal control vector $\mathbf{u}_0(t)$ is, for $T - t$ units of time remaining in the process and $\mathbf{z}(t)$ (instead of $\mathbf{x}(t)$ as before) the current state,

$$\mathbf{u}_0(t) = -\mathbf{B}^*(e^{\mathbf{A}(T-t)})^*\,\mathbf{P}\left[\lambda\mathbf{I} + \int_0^{T-t} e^{\mathbf{A}s}\mathbf{B}\mathbf{B}^*(e^{\mathbf{A}s})^*\,\mathbf{P}\,ds\right]^{-1} e^{\mathbf{A}(T-t)}\,\mathbf{z}(t) \tag{9–53}$$

where in this case the matrices $\mathbf{A}$ and $\mathbf{B}$ are given by Eq. (9–52). Equation (9–53) gives the optimal control vector $\mathbf{u}_0(t)$ for any time t where $0 \leqslant t \leqslant T$.

To find the equations for the time-varying matrices $\mathbf{F}(T - t)$ and $\mathbf{H}(T-t)$ note that

$$\mathbf{u}(t) = \boldsymbol{\eta}(t) - \mathbf{y}(t) = \mathbf{H}(T - t)\boldsymbol{\xi}(t) - \mathbf{F}(T - t)\mathbf{x}(t)$$

and in matrix form

$$\mathbf{u}(t) = [-\mathbf{F}(T - t)\,|\,\mathbf{H}(T - t)]\begin{bmatrix}\mathbf{x}(t) \\ \boldsymbol{\xi}(t)\end{bmatrix} = [-\mathbf{F}(T - t)\,|\,\mathbf{H}(T - t)]\,\mathbf{z}(t)$$

By denoting $\mathbf{F}(T - t)$ and $\mathbf{H}(T - t)$ corresponding to the optimal control vector $\mathbf{u}_0(t)$ as $\mathbf{F}_0(T - t)$ and $\mathbf{H}_0(T - t)$, we obtain

$$\mathbf{u}_0(t) = [-\mathbf{F}_0(T - t)\,|\,\mathbf{H}_0(T - t)]\,\mathbf{z}(t) \tag{9–54}$$

Equation (9–53) and Eq. (9–54) represent the same optimal control vector. Hence the matrix $[-\mathbf{F}_0(T - t)\,|\,\mathbf{H}_0(T - t)]$ can be equated to the coefficient of $\mathbf{z}(t)$ in Eq. (9–53). Therefore, $\mathbf{F}_0(T - t)$ and $\mathbf{H}_0(T - t)$ are obtained from the relationship

$$[-\mathbf{F}_0(T - t)\,|\,\mathbf{H}_0(T - t)] = -\mathbf{B}^*(e^{\mathbf{A}(T-t)})^*\,\mathbf{P}\left[\lambda\mathbf{I} + \int_0^{T-t} e^{\mathbf{A}s}\mathbf{B}\mathbf{B}^*(e^{\mathbf{A}s})^*\,\mathbf{P}\,ds\right]^{-1} e^{\mathbf{A}(T-t)} \tag{9–55}$$

By equating the corresponding elements of the matrices on both sides of the equation, we obtain the expressions for the elements of matrices $\mathbf{F}_0(T - t)$ and $\mathbf{H}_0(T - t)$ for optimal performance. Using $\mathbf{A}$ and $\mathbf{B}$ given by Eq. (9–52), we obtain

$$\begin{aligned} e^{\mathbf{A}s} &= \left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I} \end{array}\right] + s\left[\begin{array}{c|c} \mathbf{A}_x & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{A}_\xi \end{array}\right] + \frac{s^2}{2!}\left[\begin{array}{c|c} \mathbf{A}_x^2 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{A}_\xi^2 \end{array}\right] + \cdots \\ &= \left[\begin{array}{c|c} e^{\mathbf{A}_x s} & \mathbf{0} \\ \hline \mathbf{0} & e^{\mathbf{A}_\xi s} \end{array}\right] \end{aligned}$$

and

$$e^{\mathbf{A}s}\mathbf{B}\mathbf{B}^*(e^{\mathbf{A}s})^*\mathbf{P} = \left[\begin{array}{c|c} e^{\mathbf{A}_x s} & \mathbf{0} \\ \hline \mathbf{0} & e^{\mathbf{A}_t s} \end{array}\right]\left[\begin{array}{c} \mathbf{B}_x \\ \hline \mathbf{0} \end{array}\right][\mathbf{B}_x^* \,|\, \mathbf{0}]\left[\begin{array}{c|c} (e^{\mathbf{A}_x s})^* & \mathbf{0} \\ \hline \mathbf{0} & (e^{\mathbf{A}_t s})^* \end{array}\right]\left[\begin{array}{c|c} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \hline \mathbf{P}_{21} & \mathbf{P}_{22} \end{array}\right]$$

$$= \left[\begin{array}{c|c} e^{\mathbf{A}_x s}\mathbf{B}_x\mathbf{B}_x^*(e^{\mathbf{A}_x s})^* & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]\left[\begin{array}{c|c} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \hline \mathbf{P}_{21} & \mathbf{P}_{22} \end{array}\right]$$

$$= \left[\begin{array}{c|c} e^{\mathbf{A}_x s}\mathbf{B}_x\mathbf{B}_x^*(e^{\mathbf{A}_x s})^*\mathbf{P}_{11} & e^{\mathbf{A}_x s}\mathbf{B}_x\mathbf{B}_x^*(e^{\mathbf{A}_x s})^*\mathbf{P}_{12} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]$$

Define

$$\mathbf{M}_x(T-t) = \int_0^{T-t}\left[\begin{array}{c|c} e^{\mathbf{A}_x s}\mathbf{B}_x\mathbf{B}_x^*(e^{\mathbf{A}_x s})^*\mathbf{P}_{11} & e^{\mathbf{A}_x s}\mathbf{B}_x\mathbf{B}_x^*(e^{\mathbf{A}_x s})^*\mathbf{P}_{12} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]ds$$

and

$$\left[\begin{array}{c|c} \mathbf{N}_{11} & \mathbf{N}_{12} \\ \hline \mathbf{N}_{21} & \mathbf{N}_{22} \end{array}\right] = \mathbf{P}\,[\lambda\mathbf{I} + \mathbf{M}_x(T-t)]^{-1}$$

then Eq. (9–55) becomes as follows:

$$[-\mathbf{F}_0(T-t) \,|\, \mathbf{H}_0(T-t)]$$

$$= -[\mathbf{B}_x^* \,|\, \mathbf{0}]\left[\begin{array}{c|c} (e^{\mathbf{A}_x(T-t)})^* & \mathbf{0} \\ \hline \mathbf{0} & (e^{\mathbf{A}_t(T-t)})^* \end{array}\right]\left[\begin{array}{c|c} \mathbf{N}_{11} & \mathbf{N}_{12} \\ \hline \mathbf{N}_{21} & \mathbf{N}_{22} \end{array}\right]\left[\begin{array}{c|c} e^{\mathbf{A}_x(T-t)} & \mathbf{0} \\ \hline \mathbf{0} & e^{\mathbf{A}_t(T-t)} \end{array}\right]$$

$$= -[\mathbf{B}_x^*(e^{\mathbf{A}_x(T-t)})^*\mathbf{N}_{11}e^{\mathbf{A}_x(T-t)} \,|\, \mathbf{B}_x^*(e^{\mathbf{A}_x(T-t)})^*\mathbf{N}_{12}e^{\mathbf{A}_t(T-t)}]$$

Hence we obtain

$$\begin{aligned} \mathbf{F}_0(T-t) &= \mathbf{B}_x^*(e^{\mathbf{A}_x(T-t)})^*\mathbf{N}_{11}e^{\mathbf{A}_x(T-t)} \\ \mathbf{H}_0(T-t) &= -\mathbf{B}_x^*(e^{\mathbf{A}_x(T-t)})^*\mathbf{N}_{12}e^{\mathbf{A}_t(T-t)} \end{aligned} \qquad (9\text{–}56)$$

Equations (9–56) define $\mathbf{F}_0(T-t)$ and $\mathbf{H}_0(T-t)$ for optimal performance.

In order to mechanize the optimal control system, it is necessary to program the feedback time-varying matrix $\mathbf{F}_0(T-t)$ and the input time-varying matrix $\mathbf{H}_0(T-t)$ as functions of time remaining in the process, $T-t$. In other words, the time remaining, $T-t$, is a quantity that must be available throughout the control process in order to realize the optimal control law specified by the theory.

In presenting the method, we intentionally used a simple quadratic or Hermitian performance index. As a matter of course the method applies to more complicated quadratic or Hermitian performance indexes as well.

Note that the system is not designed for any particular input, but is designed for a polynomial in the general form. The elements of $\mathbf{F}_0(T-t)$ and $\mathbf{H}_0(T-t)$ can be adjusted according to the degree of the polynomial input.

As an example of the design of an optimal control system by the present method, consider a simple case where the plant is governed by the following state space equation:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}[u]$$

where x_1, x_2, and u are real quantities. We assume that the input to the system consists of a family of ramp inputs. Then, the input satisfies the following state space equation:

$$\begin{bmatrix} \dot{\xi}_1 \\ \dot{\xi}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}$$

If the performance index is given by

$$J(\mathbf{c}, T) = \mathbf{z}'(T)\mathbf{P}\mathbf{z}(T) + \lambda\int_0^T u'(t)u(t)dt$$

where

$$\mathbf{z}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \xi_1(t) \\ \xi_2(t) \end{bmatrix}, \qquad \mathbf{P} = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}$$

then the elements of $\mathbf{F}_0(T-t)$ and $\mathbf{H}_0(T-t)$ of the optimal control system are determined as follows (for the derivation of these matrices refer to Prob. A-9-7):

$$\begin{aligned} \mathbf{F}_0(T-t) &= \frac{1}{\Delta_1}[\lambda(T-t) + \tfrac{1}{2}(T-t)^2 \,|\, \lambda + \lambda(T-t)^2 + \tfrac{1}{3}(T-t)^3] \\ &= [f_{11}(T-t) \,|\, f_{12}(T-t)] \\ \mathbf{H}_0(T-t) &= \frac{1}{\Delta_1}[\lambda(T-t) + \tfrac{1}{2}(T-t)^2 \,|\, \lambda + \lambda(T-t)^2 + \tfrac{1}{3}(T-t)^3] \\ &= [h_{11}(T-t) \,|\, h_{12}(T-t)] \end{aligned}$$

where

$$\Delta_1 = \lambda^2 + \lambda(T-t) + \frac{\lambda}{3}(T-t)^3 + \frac{1}{12}(T-t)^4$$

In this simple case, $\mathbf{F}_0(T-t)$ and $\mathbf{H}_0(T-t)$ become the same. Figure 9–13 (A) shows a block diagram of the optimal control system. Figure 9–13 (B) is a simplified form of the block diagram of Fig. 9–13 (A). If the input to the system consists of a family of ramp inputs and if the time derivative of the input is constant for each time interval $0 \leqslant t \leqslant T$, then the system designed here is optimal for the given performance index.

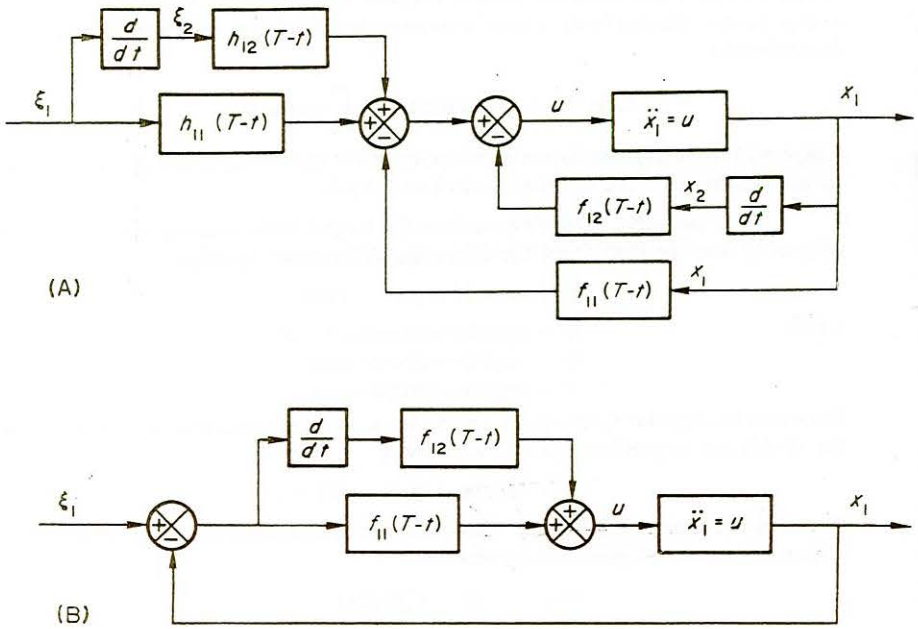


Figure 9-13 (A) Block diagram for optimal control of plant defined by $\ddot{x}_1 = u$; (B) simplified block diagram.

9-6 CONCLUDING COMMENTS

The use of a quadratic or Hermitian performance index in conjunction with a linear plant will result in a linear time-varying control system. The feedback and feedforward gains are not functions of the state of the plant but merely functions of the remaining time in the process. Note that functional forms of weighting factors appearing in performance indexes should be chosen from the consideration of the time domain performance and cost requirements.

As shown in this chapter, quadratic or Hermitian performance indexes are mathematically tractable. Synthesis procedures are straightforward. The resulting optimal controllers, however, are not simple to build in the sense that they are time-varying.

It is important to remember that before we decide whether to build the optimal control system or something inferior but simpler, we should carefully evaluate a measure of the degree to which the performance of the complex but optimal control system exceeds that of a simpler but nonoptimal control system. Unless the complex but optimal control system can be justified, we will not build extremely complicated optimal control systems. Systems

synthesized by optimal control theory should then be interpreted as a basis for defining the ultimate performance limitation under the given performance index. Once we know such ultimate performance limitation we can make efforts to design a system that is simpler yet close to optimal.

EXERCISE PROBLEMS AND SOLUTIONS

Problem A-9-1 (9–8) Suppose that a linear system is described by

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u}$$

where
$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{A}(t) = n \times n$ matrix
$\mathbf{B}(t) = n \times r$ matrix

The elements of $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are assumed to be continuous functions of time. Assume that the system is completely state controllable. It is desired to bring the state of the system from $\mathbf{c}$ at $t = t_0$ to $\mathbf{x}_f$ in time T. Obtain the optimal control vector which brings the state from $\mathbf{c}$ to $\mathbf{x}_f$ in time T with the minimum of the quantity

$$\int_{t_0}^{T+t_0} \mathbf{u}^*(t)\mathbf{u}(t)\,dt$$

Solution Any integrable control vector $\mathbf{u}(t)$ transfers the initial state $\mathbf{c}$ to the final state $\mathbf{x}_f$ if the following equation is satisfied:

$$\mathbf{x}(T + t_0) = \mathbf{x}_f = \Phi(T + t_0, t_0)\,\mathbf{c} + \int_{t_0}^{T+t_0} \Phi(T + t_0, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau \qquad (9\text{–}57)$$

where $\Phi(t, t_0)$ is the solution of the following matrix differential equation:

$$\dot{\Phi}(t, t_0) = \mathbf{A}(t)\Phi(t, t_0), \qquad \Phi(t_0, t_0) = \mathbf{I}$$

There is one control vector $\mathbf{u}(t)$ such that it also yields the minimum value for the quantity

$$\int_{t_0}^{T+t_0} \mathbf{u}^*(t)\mathbf{u}(t)\,dt$$

Denote this particular $\mathbf{u}(t)$ by $\mathbf{u}_0(t)$. We can obtain $\mathbf{u}_0(t)$ by applying calculus of variations using Eq. (9–57) as the equality constraint. Namely, first determine $\mathbf{u}(t)$ such that

$$\int_{t_0}^{T+t_0} [\mathbf{u}^*(t)\mathbf{u}(t) + \boldsymbol{\lambda}^*\Phi(T + t_0, t)\mathbf{B}(t)\mathbf{u}(t)]\,dt$$

where $\boldsymbol{\lambda}$ is a Lagrange multiplier (an n vector), is minimum and then determine $\boldsymbol{\lambda}$ by use of Eq. (9–57). The result is

$$\mathbf{u}_0(t) = \mathbf{B}^*(t)\Phi^*(T + t_0, t)\mathbf{V}^{-1}(T + t_0, t_0)\,[\mathbf{x}(T + t_0) - \Phi(T + t_0, t_0)\mathbf{c}]$$

where

$$\mathbf{V}(T + t_0, t_0) = \int_{t_0}^{T+t_0} \Phi(T + t_0, t)\mathbf{B}(t)\mathbf{B}^*(t)\Phi^*(T + t_0, t)\,dt \qquad (9\text{–}58)$$

Problem A-9-2 (9–8) Consider the system defined in Prob. A-9-1. It is desired to bring the state of the system from $\mathbf{c}$ to $\mathbf{x}_f$ in time T. As time T approaches zero the quantity

$$\int_{t_0}^{T+t_0} \|\mathbf{u}(t)\|^2 \, dt$$

approaches infinity. Therefore, there is a least time T_0 such that

$$\int_{t_0}^{T_0+t_0} \|\mathbf{u}(t)\|^2 \, dt = K \tag{9–59}$$

where K is a given positive real number.

Show that T_0 is the smallest value of T for which the following matrix is positive semidefinite:

$$K\mathbf{V}(T + t_0, t_0) - [\mathbf{x}(T + t_0) - \Phi(T + t_0, t_0)\mathbf{c}]\,[\mathbf{x}(T + t_0) - \Phi(T + t_0, t_0)\mathbf{c}]^*$$

where $\mathbf{V}(T + t_0, t_0)$ and $\Phi(t, t_0)$ are defined in Prob. A-9-1.

Solution The solution to the system differential equation is given by

$$\mathbf{x}(t) = \Phi(t, t_0)\mathbf{c} + \Phi(t, t_0) \int_{t_0}^{t} \Phi^{-1}(\tau, t_0)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau$$

$$= \Phi(t, t_0)\mathbf{c} + \int_{t_0}^{t} \Phi(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau$$

From this equation we obtain

$$\mathbf{x}(T + t_0) - \Phi(T + t_0, t_0)\mathbf{c} = \int_{t_0}^{T+t_0} \Phi(T + t_0, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau$$

Hence, if $\boldsymbol{\lambda}$ is an arbitrary n vector ($\boldsymbol{\lambda} \neq \mathbf{0}$)

$$\|\boldsymbol{\lambda}^*[\mathbf{x}(T + t_0) - \Phi(T + t_0, t_0)\mathbf{c}]\| = \left\| \int_{t_0}^{T+t_0} \boldsymbol{\lambda}^*\Phi(T + t_0, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau \right\| \tag{9–60}$$

Using the following Schwarz inequality,

$$\left\| \int_{t_0}^{t_1} \mathbf{v}^*(\tau)\mathbf{u}(\tau)\, d\tau \right\| \leqslant \left[\int_{t_0}^{t_1} \|\mathbf{v}(\tau)\|^2 d\tau \right]^{\frac{1}{2}} \left[\int_{t_0}^{t_1} \|\mathbf{u}(\tau)\|^2 d\tau \right]^{\frac{1}{2}}$$

where $\mathbf{u}(\tau)$ and $\mathbf{v}(\tau)$ are r vectors, Eq. (9–60) becomes

$$\|\boldsymbol{\lambda}^*[\mathbf{x}(T + t_0) - \Phi(T + t_0, t_0)\mathbf{c}]\| \leqslant \left[\int_{t_0}^{T+t_0} \|\boldsymbol{\lambda}^*\Phi(T + t_0, \tau)\mathbf{B}(\tau)\|^2 d\tau \int_{t_0}^{T+t_0} \|\mathbf{u}(\tau)\|^2 d\tau \right]^{1/2}$$

Squaring both sides of this last inequality and making use of Eq. (9–58) and Eq. (9–59) we obtain

$$\boldsymbol{\lambda}^* \{[\mathbf{x}(T + t_0) - \Phi(T + t_0, t_0)\mathbf{c}] \cdot [\mathbf{x}(T + t_0) - \Phi(T + t_0, t_0)\mathbf{c}]^*\} \boldsymbol{\lambda} \leqslant K\boldsymbol{\lambda}^*\mathbf{V}(T + t_0, t_0)\boldsymbol{\lambda}$$

If the specific state $\mathbf{x}(T + t_0) = \mathbf{x}_f$ is controllable relative to Eq. (9–59), this last inequality should be satisfied for any $\boldsymbol{\lambda}$ where $\boldsymbol{\lambda} \neq \mathbf{0}$. Therefore,

$$\mathbf{C}(T + t_0, t_0) = K\mathbf{V}(T + t_0, t_0) - [\mathbf{x}(T + t_0) - \Phi(T + t_0, t_0)\mathbf{c}] \cdot [\mathbf{x}(T + t_0) - \Phi(T + t_0, t_0)\mathbf{c}]^*$$

should be positive semidefinite. The minimum time which the state $\mathbf{x}_f$ can be reached is then T_0, the smallest T for which $\mathbf{C}(T + t_0, t_0)$ is positive semidefinite.

Problem A-9-3 (9–7, 9–10) Consider a tumbling satellite in orbit. The problem is to reduce the angular velocity of the rigid body rotating about its center of mass. The control process consists of applying a time-varying torque vector to the rotating body in such a manner that the following performance index is minimized:

$$J(\mathbf{c}, T) = \mathbf{x}'(T)\mathbf{x}(T) + \lambda \int_0^T \mathbf{u}'(t)\mathbf{u}(t)\, dt$$

where $\mathbf{x}(t)$ is the angular momentum vector of the rigid body and $\mathbf{u}(t)$ is the applied torque vector. Find the optimal control vector $\mathbf{u}_0(t)$.

Solution Euler's equations for a rigid body rotating about its center of gravity are obtained from the following differential equation:

$$\dot{\mathbf{H}} + \boldsymbol{\omega} \times \mathbf{H} = \mathbf{T}, \qquad \mathbf{H}(0) = \mathbf{c} \tag{9–61}$$

where

$\mathbf{H}$ = angular momentum vector
$\boldsymbol{\omega}$ = angular velocity vector
$\mathbf{T}$ = external torque vector

Denoting the angular momentum vector by $\mathbf{x}$ and the applied torque vector by $\mathbf{u}$, Eq. (9–61) can be put into the standard form

$$\dot{\mathbf{x}} = \boldsymbol{\Omega}(\mathbf{x})\mathbf{x} + \mathbf{u}, \qquad \mathbf{x}(0) = \mathbf{c}$$

Denoting the moments of inertia about the principal axes of inertia as A, B, and C respectively, this last equation becomes

$$\dot{\mathbf{x}} = \begin{bmatrix} A\dot{\omega}_1 \\ B\dot{\omega}_2 \\ C\dot{\omega}_3 \end{bmatrix} = \begin{bmatrix} (B - C)\omega_2\omega_3 \\ (C - A)\omega_3\omega_1 \\ (A - B)\omega_1\omega_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & \omega_3 & -\omega_2 \\ -\omega_3 & 0 & \omega_1 \\ \omega_2 & -\omega_1 & 0 \end{bmatrix} \begin{bmatrix} A\omega_1 \\ B\omega_2 \\ C\omega_3 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$$

or

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{C}x_3 & -\frac{1}{B}x_2 \\ -\frac{1}{C}x_3 & 0 & \frac{1}{A}x_1 \\ \frac{1}{B}x_2 & -\frac{1}{A}x_1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$$

Here x_1, x_2, and x_3 are state variables, and u_1, u_2, and u_3 are components of the applied torque about the principal axes of inertia. Note that $\boldsymbol{\Omega}(\mathbf{x})$ is a skew symmetric matrix. The minimized performance index is

$$f(\mathbf{c}, T) = \min_{\mathbf{u}} \left[\mathbf{x}'(T)\mathbf{x}(T) + \lambda \int_0^T \mathbf{u}'(t)\mathbf{u}(t)\, dt \right] \tag{9–62}$$

The partial differential equation which should be satisfied by $f(\mathbf{c}, T)$ is given by

$$\frac{\partial f}{\partial T} = \min_{\mathbf{u}}\left[\lambda \mathbf{u}'(0)\mathbf{u}(0) + (\nabla f)' \frac{d\mathbf{c}}{dt}\right] \tag{9-63}$$

where

$$(\nabla f)' = \left[\frac{\partial f}{\partial c_1} \quad \frac{\partial f}{\partial c_2} \quad \frac{\partial f}{\partial c_3}\right]$$

and

$$\frac{d\mathbf{c}}{dt} = \boldsymbol{\Omega}(\mathbf{c})\mathbf{c} + \mathbf{u}$$

The optimal torque vector $\mathbf{u}_0(0)$ is obtained as

$$\mathbf{u}_0(0) = -\frac{1}{2\lambda}\nabla f \tag{9-64}$$

Substituting Eq. (9-64) into Eq. (9-63),

$$\frac{\partial f}{\partial T} = (\nabla f)'\left(\boldsymbol{\Omega}(\mathbf{c})\mathbf{c} - \frac{1}{4\lambda}\nabla f\right) \tag{9-65}$$

From Eq. (9-62) we know that $f(\mathbf{c}, 0) = \mathbf{c}'\mathbf{c}$. Therefore, we assume a solution of the form

$$f(\mathbf{c}, T) = \mathbf{c}'\mathbf{S}(T)\mathbf{c} \tag{9-66}$$

where $\mathbf{S}(T)$ is symmetric and $\mathbf{S}(0) = \mathbf{I}$. Substituting Eq. (9-66) into Eq. (9-65) and rearranging,

$$\mathbf{c}'\left[\frac{d\mathbf{S}}{dT} - \mathbf{S}\boldsymbol{\Omega} - \boldsymbol{\Omega}'\mathbf{S} + \frac{1}{\lambda}\mathbf{S}\mathbf{S}\right]\mathbf{c} = 0$$

The solution to the following matrix Riccati equation

$$\frac{d\mathbf{S}}{dT} - \mathbf{S}\boldsymbol{\Omega} - \boldsymbol{\Omega}'\mathbf{S} + \frac{1}{\lambda}\mathbf{S}\mathbf{S} = \mathbf{0} \tag{9-67}$$

can be found by the standard technique. In this problem, however, because of the skew symmetric nature of the matrix $\boldsymbol{\Omega}$ we may assume $\mathbf{S}(T)$ in the following from:

$$\mathbf{S}(T) = s(T)\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}, \qquad s(0) = 1$$

Then,

$$\mathbf{S}\boldsymbol{\Omega} + \boldsymbol{\Omega}'\mathbf{S} = \mathbf{0}$$

and Eq. (9-67) is simplified to give

$$\frac{d\mathbf{S}}{dT} + \frac{1}{\lambda}\mathbf{S}\mathbf{S} = \mathbf{0}$$

$s(T)$ is then determined from

$$\frac{ds(T)}{dT} + \frac{1}{\lambda}s^2(T) = 0$$

which can be solved easily to give

$$s(T) = \frac{\lambda}{\lambda + T}$$

Hence,

$$f(\mathbf{c}, T) = \mathbf{c}'\left(\frac{\lambda}{\lambda + T}\mathbf{I}\right)\mathbf{c} = \frac{\lambda}{\lambda + T}\mathbf{c}'\mathbf{c}$$

The optimal control vector is

$$\mathbf{u}_0(0) = -\frac{1}{\lambda + T}\mathbf{c}$$

Letting $T - t$ be the duration of time remaining in the process and $\mathbf{x}(t)$ be the current state, the optimal control vector $\mathbf{u}_0(t)$ becomes as follows:

$$\mathbf{u}_0(t) = -\frac{1}{\lambda + (T - t)}\mathbf{x}(t)$$

Problem A-9-4 Consider the control system shown in Fig. 9–14. The performance index is given by

$$J(\mathbf{c}, T) = \mathbf{x}'(T)\mathbf{x}(T) + \lambda\int_0^T u'(t)u(t)\,dt$$

where

$$\mathbf{x} = \begin{bmatrix}x_1\\ x_2\end{bmatrix} = \begin{bmatrix}x\\ \dot{x}\end{bmatrix}, \qquad \mathbf{c} = \mathbf{x}(0)$$

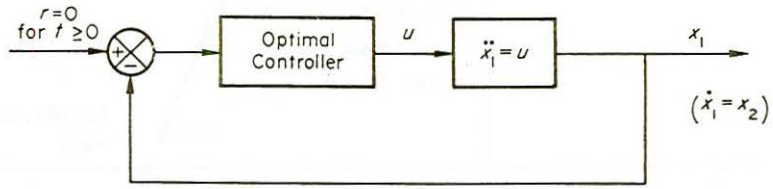


Figure 9-14 Control system.

and λ is a positive constant. The optimal control function u_0 which minimizes this performance index can be obtained in the following form (See Example 9–1.):

$$u_0(t) = -\frac{\{\lambda(T-t) + \frac{1}{2}(T-t)^2\}x_1(t) + \{\lambda + \lambda(T-t)^2 + \frac{1}{3}(T-t)^3\}x_2(t)}{\lambda^2 + \lambda(T-t) + \frac{1}{3}\lambda(T-t)^3 + \frac{1}{12}(T-t)^4}$$

Assuming $\lambda = 1$, $T = 1$ and the initial condition $x_1(0) = 2$, $x_2(0) = 2$, draw a state plane trajectory for the time interval $0 \leqslant t \leqslant T$.

Solution Divide T into N equal time increments Δt, or $N\Delta t = T$, and then compute $u_0(n\,\Delta t)$, where, for $\lambda = 1$,

$$u_0(n\Delta t) = -\frac{[(T - n\Delta t) + \frac{1}{2}(T - n\Delta t)^2]\, x_1(n\Delta t)}{1 + (T - n\Delta t) + \frac{1}{3}(T - n\Delta t)^3 + \frac{1}{12}(T - n\Delta t)^4}$$
$$- \frac{[1 + (T - n\Delta t)^2 + \frac{1}{3}(T - n\Delta t)^3]\, x_2(n\Delta t)}{1 + (T - n\Delta t) + \frac{1}{3}(T - n\Delta t)^3 + \frac{1}{12}(T - n\Delta t)^4}$$

If we divide, for example, T into five equal time increments, then $\Delta t = 0.2$, and

$$\begin{aligned}
u_0(0) &= -0.62x_1(0) - 0.96x_2(0)\\
u_0(0.2) &= -0.56x_1(0.2) - 0.90x_2(0.2)\\
u_0(0.4) &= -0.46x_1(0.4) - 0.85x_2(0.4)\\
u_0(0.6) &= -0.34x_1(0.6) - 0.83x_2(0.6)\\
u_0(0.8) &= -0.18x_1(0.8) - 0.87x_2(0.8)\\
u_0(1.0) &= -x_2(1.0)
\end{aligned}$$

The state plane trajectory can be plotted by the use of the equation

$$\frac{dx_2}{dx_1} = \frac{u}{x_2} = \text{slope of trajectory}$$

The value of u for each time interval, 0.2 sec, is assumed to be constant. (The values of u are taken at the midpoint of each interval.) Figure 9–15 shows the state plane trajectory.

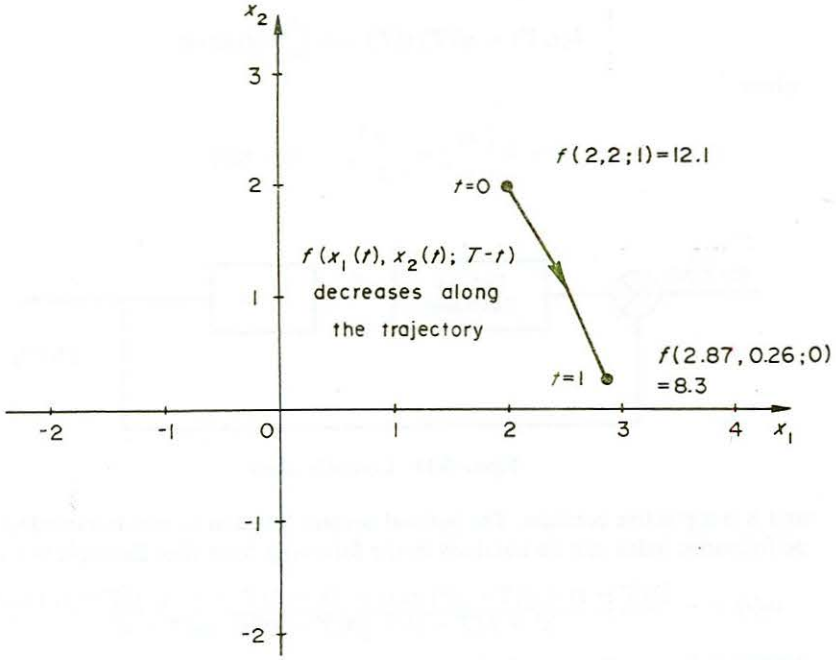


Figure 9-15 State plane trajectory.

Referring to Example 9–1, $f(\mathbf{c}, T)$ for this problem can be given by

$$f(\mathbf{c}, T) = \frac{\lambda\{(\lambda + T)c_1^2 + 2[\lambda T + (T^2/2)]c_1c_2 + [\lambda + \lambda T^2 + (T^3/3)]c_2^2\}}{\lambda^2 + \lambda T + \lambda(T^3/3) + (T^4/12)}$$

Substitution of $\lambda = 1$ and $T = 1$ into $f(\mathbf{c}, T)$ yields

$$f(c_1, c_2; 1) = 0.83c_1^2 + 1.24c_1c_2 + 0.96c_2^2$$

At the starting point, $c_1 = 2$ and $c_2 = 2$. Hence

$$f(2, 2; 1) = 12.1$$

At the end of the process,

$$f(x_1(1), x_2(1); 0) = x_1^2(1) + x_2^2(1), \qquad x_1(1) = 2.87, \qquad x_2(1) = 0.26$$

Hence

$$f(2.87,\ 0.26;\ 0) = 8.3$$

By a graphical computation

$$\int_0^1 u_0'(t)u_0(t)\, dt$$

becomes equal to 3.8. Hence, we can verify that

$$f(2, 2; 1) = f(2.87, 0.26; 0) + \int_0^1 u_0'(t)u_0(t)\, dt$$

Problem A-9-5 Given the system configuration shown in Fig. 9–14 and the performance index

$$J(\mathbf{c}, T) = \mathbf{x}'(T)\mathbf{x}(T) + \int_0^T \mathbf{x}'(t)\mathbf{x}(t)\, dt + \int_0^T u'(t)u(t)\, dt$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ \dot{x}_1 \end{bmatrix}, \qquad \mathbf{c} = \mathbf{x}(0)$$

obtain the optimal control function $u_0(t)$ and the minimized performance index $f(\mathbf{c}, T) = \min_u J(\mathbf{c}, T)$. Plot a few curves of constant $f(\mathbf{c}, T)$ for $T = 1$.

Solution The state space equation of the system is given by

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

From Eq. (9–38) and Eq. (9–39) the solutions for the optimal control function and the minimized performance index for this problem are given (with $\lambda = 1$) respectively

$$u_0(t) = -\mathbf{B}'[\Phi_{21}(T - t) + \Phi_{22}(T - t)]\,[\Phi_{11}(T - t) + \Phi_{12}(T - t)]^{-1}\mathbf{x}(t)$$

$$f(\mathbf{c}, T) = \mathbf{c}'[\Phi_{21}(T) + \Phi_{22}(T)]\,[\Phi_{11}(T) + \Phi_{12}(T)]^{-1}\mathbf{c}$$

where the $\Phi_{ij} \equiv \Phi_{ij}(T)$ $(i, j = 1, 2)$ are defined by the equations

$$e^{\mathbf{M}T} = \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix}, \qquad \mathbf{M} = \begin{bmatrix} -\mathbf{A} & \mathbf{BB}' \\ \mathbf{I} & \mathbf{A}' \end{bmatrix}$$

For this problem

$$\mathbf{M} = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

For the following 4×4 matrix $\mathbf{G}$:

$$\mathbf{G} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

$e^{\mathbf{G}T}$ was given in Prob. A-6-12 as follows:

$$e^{\mathbf{G}T} = \begin{bmatrix} pb - qa & -pa + qb & 2qb & -2qa \\ pa + qb & pb + qa & 2qa & 2qb \\ -pa + qb & -2qa & pb - qa & -pa - qb \\ 2qa & pa + qb & pa - qb & pb + qa \end{bmatrix} \tag{9-68}$$

where

$$p = \frac{1}{2}\left(e^{(\sqrt{3}/2)T} + e^{-(\sqrt{3}/2)T}\right)$$

$$q = \frac{1}{2\sqrt{3}}\left(e^{(\sqrt{3}/2)T} - e^{-(\sqrt{3}/2)T}\right)$$

$$a = \sin\frac{T}{2}$$

$$b = \cos\frac{T}{2}$$

Notice that since

$$\mathbf{G} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \mathbf{M} \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}$$

where $\mathbf{I}$ is the 2×2 identity matrix, we have

$$e^{\mathbf{G}T} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} e^{\mathbf{M}T} \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \Phi_{22} & \Phi_{21} \\ \Phi_{12} & \Phi_{11} \end{bmatrix}$$

From the definitions of Φ_{21} and Φ_{22} and from Eq. (9–68)

$$\Phi_{21} + \Phi_{22} = \begin{bmatrix} pb - qa + 2qb & -pa + qb - 2qa \\ pa + qb + 2qa & pb + qa + 2qb \end{bmatrix}$$

Similarly

$$\Phi_{11} + \Phi_{12} = \begin{bmatrix} -pa + qb + pb - qa & -2qa - pa - qb \\ 2qa + pa - qb & pa + qb + pb + qa \end{bmatrix}$$

Inverting

$$[\Phi_{11} + \Phi_{12}]^{-1} = \frac{1}{\Delta}\begin{bmatrix} pa + qb + pb + qa & 2qa + pa + qb \\ -2qa - pa + qb & -pa + qb + pb - qa \end{bmatrix}$$

where $\Delta = 3q^2a^2 + 2pq + p^2b^2$

The optimal control function is obtained from

$$u_0(t) = -[0 \quad 1]\,[\Phi_{21}(T-t) + \Phi_{22}(T-t)]\,[\Phi_{11}(T-t) + \Phi_{12}(T-t)]^{-1}\mathbf{x}(t)$$

By defining

$$p_1 = \frac{1}{2}\left(e^{(\sqrt{3}/2)(T-t)} + e^{-(\sqrt{3}/2)(T-t)}\right)$$

$$q_1 = \frac{1}{2\sqrt{3}}\left(e^{(\sqrt{3}/2)(T-t)} - e^{-(\sqrt{3}/2)(T-t)}\right)$$

$$a_1 = \sin\frac{T-t}{2}$$

$$b_1 = \cos\frac{T-t}{2}$$

we obtain

$$u_0(t) = \frac{-1}{\Delta}[0 \quad 1]\begin{bmatrix} p_1b_1 - q_1a_1 + 2q_1b_1 & -p_1a_1 + q_1b_1 - 2q_1a_1 \\ p_1a_1 + q_1b_1 + 2q_1a_1 & p_1b_1 + q_1a_1 + 2q_1b_1 \end{bmatrix} \cdot \begin{bmatrix} p_1a_1 + q_1b_1 + p_1b_1 + q_1a_1 & 2q_1a_1 + p_1a_1 + q_1b_1 \\ -2q_1a_1 - p_1a_1 + q_1b_1 & -p_1q_1 + q_1b_1 + p_1b_1 - q_1a_1 \end{bmatrix}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \tag{9-69}$$

The minimized performance index is given by

$$\begin{aligned} f(\mathbf{c}, T) &= \mathbf{c}'\,[\Phi_{21}(T) + \Phi_{22}(T)]\,[\Phi_{11}(T) + \Phi_{12}(T)]^{-1}\mathbf{c} \\ &= \frac{1}{\Delta}[c_1 \quad c_2]\begin{bmatrix} pb - qa + 2qb & -pa + qb - 2qa \\ pa + qb + 2qa & pb + qa + 2qb \end{bmatrix} \\ &\quad \cdot \begin{bmatrix} pa + qb + pb + qa & 2qa + pa + qb \\ -2qa - pa + qb & -pa + qb + pb - qa \end{bmatrix}\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \end{aligned} \tag{9-70}$$

For $T = 1$ we have

$$\begin{aligned} p &= \frac{1}{2}\left(e^{\sqrt{3}/2} + e^{-\sqrt{3}/2}\right) = 1.40 \\ q &= \frac{1}{2\sqrt{3}}\left(e^{\sqrt{3}/2} - e^{-\sqrt{3}/2}\right) = 0.567 \\ a &= \sin\frac{1}{2} = 0.480 \\ b &= \cos\frac{1}{2} = 0.878 \end{aligned} \tag{9-71}$$

Then Eq. (9–69) becomes

$$\begin{aligned} u_0(0) &= \frac{-1}{3.32}[0 \quad 1]\begin{bmatrix} 1.95 & -0.716 \\ 1.71 & 2.50 \end{bmatrix}\begin{bmatrix} 2.67 & 1.71 \\ -0.716 & 0.786 \end{bmatrix}\begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix}, \\ &= -0.835\,x_1(0) - 1.47\,x_2(0) \end{aligned}$$

This equation gives the initial value of the optimal control function. The optimal control function $u_o(t)$ for $0 \leqslant t \leqslant T$ is obtained from Eq. (9–69) which is a function of $T - t$. Substituting Eq. (9–71) into Eq. (9–70), we obtain

$$f(\mathbf{c}, 1) = \frac{1}{3.32}[c_1 \quad c_2]\begin{bmatrix}1.95 & -0.716\\ 1.71 & 2.50\end{bmatrix}\begin{bmatrix}2.67 & 1.71\\ -0.716 & 0.786\end{bmatrix}\begin{bmatrix}c_1\\ c_2\end{bmatrix}$$

$$= 1.72c_1^2 + 1.67c_1c_2 + 1.47c_2^2 \qquad (9\text{–}72)$$

For plotting the constant $f(\mathbf{c}, T)$ ellipses, the major and minor axes of the ellipses are found. The major and minor axes ($\hat{c}_1$ and $\hat{c}_2$ axes) of the constant $f(\mathbf{c}, T)$ ellipses are related to the c_1 and c_2 axes by the relationship

$$\begin{bmatrix}\hat{c}_1\\ \hat{c}_2\end{bmatrix} = \begin{bmatrix}\cos\xi & \sin\xi\\ -\sin\xi & \cos\xi\end{bmatrix}\begin{bmatrix}c_1\\ c_2\end{bmatrix}$$

Inverting

$$\begin{bmatrix}c_1\\ c_2\end{bmatrix} = \begin{bmatrix}\cos\xi & -\sin\xi\\ \sin\xi & \cos\xi\end{bmatrix}\begin{bmatrix}\hat{c}_1\\ \hat{c}_2\end{bmatrix}$$

Substituting

$$c_1 = \hat{c}_1\cos\xi - \hat{c}_2\sin\xi$$
$$c_2 = \hat{c}_1\sin\xi + \hat{c}_2\cos\xi$$

into Eq. (9–72),

$$f(\mathbf{c}, 1) = 1.72(\hat{c}_1\cos\xi - \hat{c}_2\sin\xi)^2 + 1.67(\hat{c}_1\cos\xi - \hat{c}_2\sin\xi)(\hat{c}_1\sin\xi + \hat{c}_2\cos\xi) + 1.47(\hat{c}_1\sin\xi + \hat{c}_2\cos\xi)^2$$

Making the coefficient of $\hat{c}_1\hat{c}_2$ equal to zero

$$-0.50\cos\xi\sin\xi + 1.67(\cos^2\xi - \sin^2\xi) = 0$$

Thus

$$\xi = 40.8° \quad \text{or} \quad -49.2°$$

A few constant $f(\mathbf{c}, 1)$ ellipses are plotted in Fig. 9–16.

Problem A-9-6 (9–6, 9–9) In general, it is not possible to design a control system for optimal performance in response to all kinds of system input signal $r(t)$. In order to avoid this difficulty, it is common to optimize the control system only with respect to a restricted class of input signals.

Let $\{r(t)\}$ be a class of system inputs. An element $r(t)$ of this class is defined as

$$r(t) = y_1(t) \qquad t \geqslant 0$$
$$= 0 \qquad t < 0$$

where $y_1(t)$ is the first element of the input state vector $\mathbf{y}(t)$, where $\mathbf{y}(t)$ is the solution of

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}$$

Namely,

$$\mathbf{y}(t) = \Phi(t)\mathbf{y}(0)$$

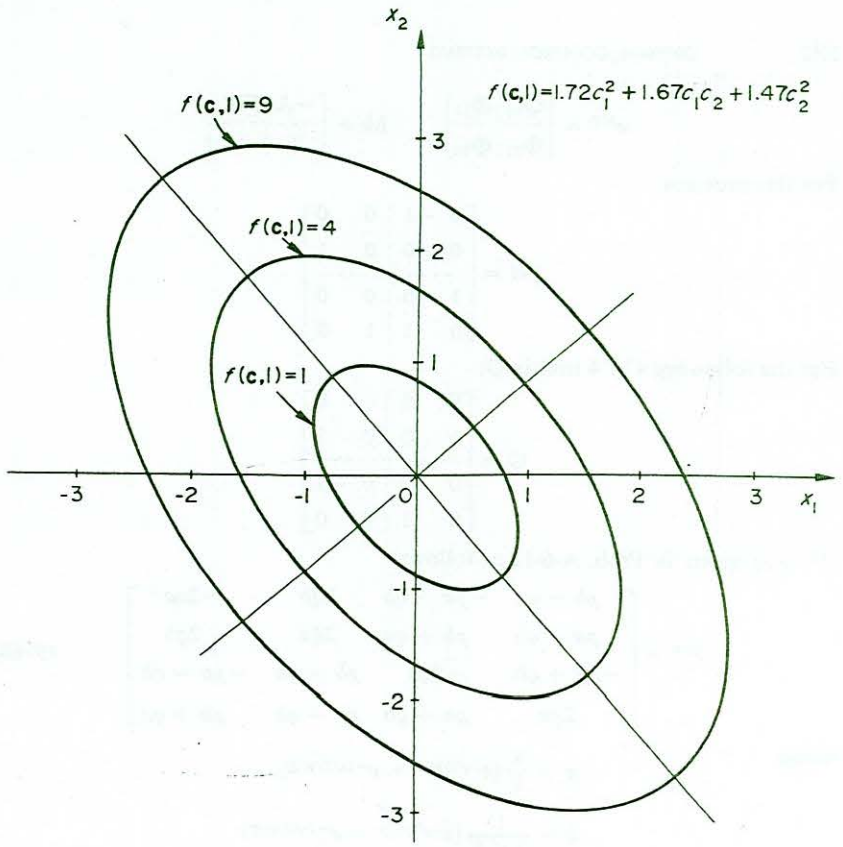


Figure 9-16 Constant $f(\mathbf{c}, 1)$ ellipses.

$\Phi(t)$ is the transition matrix of the input state space equation and $\mathbf{y}(0)$ is an arbitrary constant vector. Namely, $\Phi(t)$ determines the class of input signals for which the system is to be optimized and $\mathbf{y}(0)$ determines a particular member of that class.

Obtain $\Phi(t)$ for the following classes of inputs:

(1) The class of all step functions.
(2) The class of all step and ramp functions.
(3) The class of all sinusoidal functions having the frequency ω.

Solution

(1) The class of all step functions $\{r(t)\}$ can be described by

$$\dot{r}(t) = 0, \qquad t \neq 0$$

or putting $r = y$ we have

$$\dot{y} = Ay = 0$$

Hence, $A = 0$ and we obtain

$$\Phi(t) = e^{At} = 1$$

(2) The class of all step and ramp functions $\{r(t)\}$ can be described by

$$\ddot{r}(t) = 0, \qquad t \neq 0$$

Select the input state vector as

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} r \\ \dot{r} \end{bmatrix}$$

then the state space equation describing the input can be written as

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

The transition matrix is

$$\Phi(t) = e^{\left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right]t} = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}$$

Thus, the input is described by

$$\begin{bmatrix} r(t) \\ \dot{r}(t) \end{bmatrix} = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} r(0) \\ \dot{r}(0) \end{bmatrix}$$

(3) The class of all sinusoidal functions having the frequency ω must satisfy the differential equation

$$\ddot{r} + \omega^2 r = 0$$

Select the input state vector as

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} r \\ \dot{r} \end{bmatrix}$$

then the state space equation describing the input can be written as

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega^2 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

The transition matrix is

$$\Phi(t) = e^{\left[\begin{smallmatrix} 0 & 1 \\ -\omega^2 & 0 \end{smallmatrix}\right]t} = \begin{bmatrix} \cos \omega t & \frac{1}{\omega} \sin \omega t \\ -\omega \sin \omega t & \cos \omega t \end{bmatrix}$$

Hence

$$\begin{bmatrix} r(t) \\ \dot{r}(t) \end{bmatrix} = \begin{bmatrix} \cos \omega t & \frac{1}{\omega} \sin \omega t \\ -\omega \sin \omega t & \cos \omega t \end{bmatrix} \begin{bmatrix} r(0) \\ \dot{r}(0) \end{bmatrix}$$

Problem A-9-7 Consider the control system shown in Fig. 9–17. The equation of the plant is given by

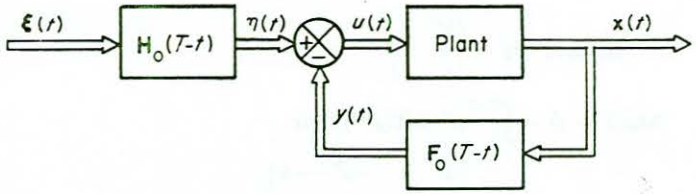


Figure 9-17 Control system.

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} [u]$$

The input $\xi_1(t)$ is assumed to be a ramp input. Thus it may be written as

$$\begin{bmatrix} \dot{\xi}_1 \\ \dot{\xi}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}, \qquad \boldsymbol{\xi} = \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}$$

The performance index is given by

$$J(\mathbf{c}, T) = \mathbf{z}'(T)\mathbf{P}\mathbf{z}(T) + \lambda \int_0^T u'(t)u(t)\,dt$$

where

$$\mathbf{c} = \mathbf{x}(0), \qquad \mathbf{z}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \xi_1(t) \\ \xi_2(t) \end{bmatrix}, \qquad \mathbf{P} = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}$$

Determine the elements of matrices $\mathbf{F}_0(T - t)$ and $\mathbf{H}_0(T - t)$ so that the system behavior minimizes the performance index.

Solution From the analysis made in Section 9–5, $\mathbf{F}_0(T - t)$ and $\mathbf{H}_0(T - t)$ can be given by the expression

$$[-\mathbf{F}_0(T-t) \;\; \mathbf{H}_0(T-t)] = -\mathbf{B}'(e^{\mathbf{A}(T-t)})'\mathbf{P} \cdot [\lambda \mathbf{I} + \int_0^{T-t} e^{\mathbf{A}s}\mathbf{B}\mathbf{B}'(e^{\mathbf{A}s})'\mathbf{P}\,ds]^{-1} e^{\mathbf{A}(T-t)}$$

where

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{A}_x & 0 \\ \hline 0 & \mathbf{A}_\xi \end{array}\right], \qquad \mathbf{B} = \left[\begin{array}{c} \mathbf{B}_x \\ \hline 0 \end{array}\right]$$

and $\mathbf{A}_x$, $\mathbf{A}_\xi$, and $\mathbf{B}_x$ are given by

$$\dot{\mathbf{x}} = \mathbf{A}_x\mathbf{x} + \mathbf{B}_x u$$
$$\dot{\boldsymbol{\xi}} = \mathbf{A}_\xi \boldsymbol{\xi}$$

Since for this problem

$$\mathbf{A}_x = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad \mathbf{B}_x = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad \mathbf{A}_\xi = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

we find

$$e^{\mathbf{A}_x s} = \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix}, \qquad e^{\mathbf{A}_t s} = \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix}$$

Hence

$$e^{\mathbf{A}_x s}\mathbf{B}_x\mathbf{B}_x'(e^{\mathbf{A}_x s})' = \begin{bmatrix} s^2 & s \\ s & 1 \end{bmatrix}$$

$$\mathbf{M}_x(T-t) = \int_0^{T-t} e^{\mathbf{A}s}\mathbf{B}\mathbf{B}'(e^{\mathbf{A}s})'\mathbf{P}\,ds$$

$$= \int_0^{T-t} \begin{bmatrix} s^2 & s & -s^2 & -s \\ s & 1 & -s & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} ds$$

$$= \begin{bmatrix} \frac{(T-t)^3}{3} & \frac{(T-t)^2}{2} & -\frac{(T-t)^3}{3} & -\frac{(T-t)^2}{2} \\ \frac{(T-t)^2}{2} & T-t & -\frac{(T-t)^2}{2} & -(T-t) \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

and

$$[\lambda\mathbf{I} + \mathbf{M}_x(T-t)]$$

$$= \begin{bmatrix} \lambda + \frac{(T-t)^3}{3} & \frac{(T-t)^2}{2} & -\frac{(T-t)^3}{3} & -\frac{(T-t)^2}{2} \\ \frac{(T-t)^2}{2} & \lambda + (T-t) & -\frac{(T-t)^2}{2} & -(T-t) \\ 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & \lambda \end{bmatrix} \tag{9-73}$$

In Prob. A-2-7 the matrix given by Eq. (9–73) was inverted. Its inverted form is

$$[\lambda\mathbf{I} + \mathbf{M}_x(T-t)]^{-1}$$

$$= \frac{1}{\Delta}\begin{bmatrix} \lambda^2(\lambda+T-t) & -\frac{\lambda^2}{2}(T-t)^2 & \lambda\left(\lambda\frac{(T-t)^3}{3} + \frac{(T-t)^4}{12}\right) & \frac{\lambda^2}{2}(T-t)^2 \\ -\frac{\lambda^2}{2}(T-t)^2 & \lambda^2\left(\lambda + \frac{(T-t)^3}{3}\right) & \frac{\lambda^2}{2}(T-t)^2 & \frac{\lambda}{12}[12\lambda(T-t) + (T-t)^4] \\ 0 & 0 & \frac{\lambda}{12}[12\lambda^2 + 12\lambda(T-t) + 4\lambda(T-t)^3 + (T-t)^4] & 0 \\ 0 & 0 & 0 & \frac{\lambda}{12}[12\lambda^2 + 12\lambda(T-t) + 4\lambda(T-t)^3 + (T-t)^4] \end{bmatrix}$$

where

$$\Delta = \lambda^4 + \lambda^3(T-t) + \frac{\lambda^3}{3}(T-t)^3 + \frac{\lambda^2}{12}(T-t)^4$$

Define

$$\mathbf{P}[\lambda\mathbf{I} + \mathbf{M}_x(T-t)]^{-1} = \left[\begin{array}{c|c} \mathbf{N}_{11} & \mathbf{N}_{12} \\ \hline \mathbf{N}_{21} & \mathbf{N}_{22} \end{array}\right]$$

Then we obtain

$$\mathbf{N}_{11} = \frac{1}{\Delta}\begin{bmatrix} \lambda^2(\lambda + T - t) & -\frac{1}{2}\lambda^2(T-t)^2 \\ -\frac{1}{2}\lambda^2(T-t)^2 & \lambda^2\left(\lambda + \frac{(T-t)^3}{3}\right) \end{bmatrix}$$

and

$$\mathbf{N}_{12} = \frac{1}{\Delta}\begin{bmatrix} -\lambda^2(\lambda + T - t) & \frac{1}{2}\lambda^2(T-t)^2 \\ \frac{1}{2}\lambda^2(T-t)^2 & -\lambda^2\left(\lambda + \frac{(T-t)^3}{3}\right) \end{bmatrix}$$

Hence

$$\mathbf{F}_0(T-t) = \mathbf{B}_x'(e^{\mathbf{A}_x(T-t)})'\mathbf{N}_{11}e^{\mathbf{A}_x(T-t)}$$

$$= \frac{1}{\Delta}[0 \quad 1]\begin{bmatrix} 1 & 0 \\ T-t & 1 \end{bmatrix}\begin{bmatrix} \lambda^2(\lambda+T-t) & -\frac{1}{2}\lambda^2(T-t)^2 \\ -\frac{1}{2}\lambda^2(T-t)^2 & \lambda^2\left(\lambda + \frac{(T-t)^3}{3}\right) \end{bmatrix}\begin{bmatrix} 1 & T-t \\ 0 & 1 \end{bmatrix}$$

$$= \frac{1}{\Delta_1}\left[\lambda(T-t) + \frac{1}{2}(T-t)^2 \;\middle|\; \lambda + \lambda(T-t)^2 + \frac{1}{3}(T-t)^3\right]$$

where $\Delta_1 = \dfrac{\Delta}{\lambda^2} = \lambda^2 + \lambda(T-t) + \dfrac{\lambda}{3}(T-t)^3 + \dfrac{1}{12}(T-t)^4$

and

$$\mathbf{H}_0(T-t) = -\mathbf{B}_x'(e^{\mathbf{A}_x(T-t)})'\mathbf{N}_{12}e^{\mathbf{A}_t(T-t)}$$

$$= \frac{1}{\Delta_1}\left[\lambda(T-t) + \frac{1}{2}(T-t)^2 \;\middle|\; \lambda + \lambda(T-t)^2 + \frac{1}{3}(T-t)^3\right]$$

In this problem $\mathbf{F}_0(T-t)$ and $\mathbf{H}_0(T-t)$ are identical. The use of the time-varying matrices $\mathbf{F}_0(T-t)$ and $\mathbf{H}_0(T-t)$ just obtained will result in the optimal performance under the given performance index.

Problem A-9-8 Consider the system given by the following state space equation:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \qquad \mathbf{x}(0) = \mathbf{c}$$

where $\mathbf{x} = n$ vector (state vector, actual state of the system)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{A} = n \times n$ constant matrix
$\mathbf{B} = n \times r$ constant matrix

The desired state of the system is a function of time and is given by $\mathbf{x}_d(t)$ where

$$\mathbf{x}_d(t) = \begin{bmatrix} x_{1d}(t) \\ x_{2d}(t) \\ \vdots \\ x_{nd}(t) \end{bmatrix}, \qquad \mathbf{x}_d(0) = \mathbf{c}_d$$

We assume that the desired state $\mathbf{x}_d(t)$ of the system is known. We further assume that $\mathbf{x}$, $\mathbf{x}_d$, $\mathbf{u}$, $\mathbf{A}$, and $\mathbf{B}$ are real. The system error is defined by the difference between the desired state and the instantaneous actual state during the control process. Thus, the system error vector is given by $\mathbf{x}_d(t) - \mathbf{x}(t)$. The performance index is given by

$$J(\mathbf{c}, T) = \int_0^T [\mathbf{x}_d(t) - \mathbf{x}(t)]' \mathbf{Q}(t) [\mathbf{x}_d(t) - \mathbf{x}(t)] dt + \int_0^T \mathbf{u}'(t) \mathbf{R}(t) \mathbf{u}(t) dt$$

where $\mathbf{Q}(t)$ = positive definite (or positive semidefinite) real symmetric matrix

$\mathbf{R}(t)$ = positive definite real symmetric matrix

The matrix $\mathbf{Q}(t)$ determines how the instantaneous errors are to be weighted during the control process. The matrix $\mathbf{R}(t)$ represents the weight of control cost with respect to the minimizing errors.

Define $f(\mathbf{c}, T)$ as the minimized performance index, or

$$f(\mathbf{c}, T) = \min_{\mathbf{u}} J(\mathbf{c}, T)$$
$$= \min_{\mathbf{u}} \left\{ \int_0^T [\mathbf{x}_d(t) - \mathbf{x}(t)]' \mathbf{Q}(t) [\mathbf{x}_d(t) - \mathbf{x}(t)] \, dt + \int_0^T \mathbf{u}'(t) \mathbf{R}(t) \mathbf{u}(t) dt \right\}$$

Show that $f(\mathbf{c}, T)$ can be determined by expanding $f(\mathbf{c}, T)$ in the following form:

$$f(\mathbf{c}, T) = k(T) - 2 \sum_{i=1}^{n} k_i(T) c_i + \sum_{i=1}^{n} \sum_{j=1}^{n} k_{ij}(T) c_i c_j$$

where $k_{ij}(T) = k_{ji}(T)$

and determining coefficients $k(T)$, $k_i(T)$, and $k_{ij}(T)$. Discuss methods of determining these coefficients.

Solution We shall first obtain the partial differential equation which should be satisfied by $f(\mathbf{c}, T)$. It is of the same form as Eq. (9–29) and can be obtained as follows:

$$\frac{\partial f}{\partial T} = \min_{\mathbf{u}} \{ [\mathbf{x}_d(0) - \mathbf{x}(0)]' \mathbf{Q}(0) [\mathbf{x}_d(0) - \mathbf{x}(0)]$$
$$+ \mathbf{u}'(0) \mathbf{R}(0) \mathbf{u}(0) + (\nabla f)' [\mathbf{A}\mathbf{x}(0) + \mathbf{B}\mathbf{u}(0)] \} \tag{9–74}$$

The value of $\mathbf{u}(0)$ that minimizes the right-hand side of Eq. (9–74) can be obtained by solving

$$\mathbf{0} = 2\mathbf{R}(0)\mathbf{u}(0) + \mathbf{B}' \nabla f$$

Hence, the optimal control vector is obtained as

$$\mathbf{u}_0(0) = -\frac{\mathbf{R}^{-1}(0)}{2} \mathbf{B}' \nabla f \tag{9–75}$$

where

$$\nabla f = \operatorname{grad} f = \begin{bmatrix} \frac{\partial f}{\partial c_1} \\ \vdots \\ \frac{\partial f}{\partial c_n} \end{bmatrix}$$

Substituting Eq. (9–75) into Eq. (9–74),

$$\frac{\partial f}{\partial T} = [\mathbf{c}_d - \mathbf{c}]' \mathbf{Q}(0) [\mathbf{c}_d - \mathbf{c}] + (\nabla f)' \mathbf{A}\mathbf{c} - \frac{1}{4} (\nabla f)' \mathbf{B} \mathbf{R}^{-1}(0) \mathbf{B}' \nabla f \tag{9–76}$$

This is the partial differential equation which $f(\mathbf{c}, T)$ must satisfy. Let us assume the following quadratic form of the solution:

$$f(\mathbf{x}(t), T - t) = [\mathbf{x}(t) - \mathbf{k}_c(T - t)]' \mathbf{S}(T - t) [\mathbf{x}(t) - \mathbf{k}_c(T - t)]$$

or

$$f(\mathbf{c}, T) = [\mathbf{c} - \mathbf{k}_c(T)]' \mathbf{S}(T) [\mathbf{c} - \mathbf{k}_c(T)]$$

By expanding $f(\mathbf{c}\, T)$, we obtain

$$f(\mathbf{c}, T) = k(T) - 2 \sum_{i=1}^{n} k_i(T) c_i + \sum_{i=1}^{n} \sum_{j=1}^{n} k_{ij}(T) c_i c_j \tag{9–77}$$

where $k_{ij}(T) = k_{ji}(T)$

This expression is called a *parametric expansion* (*9–11*) because the coefficients $k_i(T)$ and $k_{ij}(T)$ turn out to be parameters of the optimal control system. (Refer to Prob. A-9-9.) By differentiating Eq. (9–77) to obtain the derivatives of $f(\mathbf{c}. T)$ appearing in Eq. (9–76) and then by substituting these derivatives into Eq. (9–76), the terms are collected such that the resulting expression is of the following form:

$$P_0 + P_1 c_1 + P_2 c_2 + \cdots + P_n c_n + P_{11} c_1^2 + P_{12} c_1 c_2 + \cdots + P_{1n} c_1 c_n$$
$$+ P_{22} c_2^2 + P_{23} c_2 c_3 + \cdots + P_{2n} c_2 c_n + \cdots + P_{nn} c_n^2 = 0 \tag{9–78}$$

The P coefficients consist of terms containing the k parameters and their first derivatives, the desired state variables, the elements of $\mathbf{A}$ and $\mathbf{B}$ which describe the plant, and the elements of $\mathbf{Q}$ and $\mathbf{R}$, which are weighting factors. If Eq. (9–78) is to be valid for all possible values of the initial state variables, each P coefficient independently must be equal to zero. Hence,

$$P_0 = P_i = P_{ij} = 0 \qquad (i, j = 1, 2, \ldots, n) \tag{9–79}$$

Equation (9–79) yields a set of $1 + n + n(n + 1)/2$ independent ordinary first-order differential equations which define the k parameters. These may be called the k equations. As a matter of course, quantities appearing in the k equations, such as the desired state variables, matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{Q}$, and $\mathbf{R}$, must be specified before the solution can be carried out. The boundary conditions on these k parameters are

$$k(0) = k_i(0) = k_{ij}(0) = 0 \qquad (i, j = 1, 2, \ldots, n)$$

Except for simple cases, computation of $k(T - t)$, $k_i(T - t)$, $k_{ij}(T - t)$ for $0 \leqslant t \leqslant T$ will require a digital computer. Once the k parameters are determined, then the optimal control vector given by Eq. (9–75) can be written in terms of the actual state vector and the k parameters, using Eq. (9–77).

Problem A-9-9 (9–11) Consider the landing problem of an aircraft system defined by

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u, \qquad \mathbf{x}(0) = \mathbf{c}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 0 \\ 1 & 0 & 0 & 0 \\ 0 & a_{32} & a_{33} & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} b_1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

and

x_1 = aircraft pitch rate
x_2 = aircraft pitch angle
x_3 = time derivative of aircraft altitude
x_4 = aircraft altitude
u = aircraft elevator deflection

and a_{11}, a_{12}, a_{13}, a_{32}, a_{33}, and b_1 are constants. The performance index is given by

$$J(\mathbf{c}, T) = \int_0^T [\mathbf{x}_d(t) - \mathbf{x}(t)]' \mathbf{Q}(t)[\mathbf{x}_d(t) - \mathbf{x}(t)]dt + \int_0^T u'(t)u(t)dt$$

where

$$\mathbf{x}_d(t) = \begin{bmatrix} x_{1d}(t) \\ x_{2d}(t) \\ x_{3d}(t) \\ x_{4d}(t) \end{bmatrix} = \text{desired state vector (given for } 0 \leqslant t \leqslant T)$$

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{bmatrix} = \text{actual state vector}$$

$$\mathbf{Q}(t) = \begin{bmatrix} q_1(t) & 0 & 0 & 0 \\ 0 & q_2(t) & 0 & 0 \\ 0 & 0 & q_3(t) & 0 \\ 0 & 0 & 0 & q_4(t) \end{bmatrix}$$

The elements $q_1(t)$, $q_2(t)$, $q_3(t)$, and $q_4(t)$ in $\mathbf{Q}(t)$ are time-varying weighting factors which indicate the relative importance of the various terms in the performance index. (In carrying out the design of the optimal control system, these weighting factors must be selected such that the performance requirements and constraints are satisfied.)

Show that the optimal control law can be written as follows:

$$u_0(t) = b_1[k_1(T-t) - k_{11}(T-t)x_1(t) - k_{12}(T-t)x_2(t) - k_{13}(T-t)x_3(t) - k_{14}(T-t)x_4(t)]$$

Solution Define

$$f(\mathbf{c}, T) = \min_u J(\mathbf{c}, T)$$

$$= \min_u \left\{ \int_0^T [\mathbf{x}_d(t) - \mathbf{x}(t)]' \mathbf{Q}(t)[\mathbf{x}_d(t) - \mathbf{x}(t)]dt + \int_0^T u'(t)u(t)\,dt \right\}$$

The partial differential equation which should be satisfied by $f(\mathbf{c}, T)$ can be obtained as follows: [The x_i appearing in partial derivatives are $x_i(0)$.]

$$\frac{\partial f}{\partial T} = \min_u \Big\{ q_1(0)[x_{1d}(0) - x_1(0)]^2 + q_2(0)[x_{2d}(0) - x_2(0)]^2 + q_3(0)[x_{3d}(0) - x_3(0)]^2 + q_4(0)[x_{4d}(0) - x_4(0)]^2 + u(0)^2 + [a_{11}x_1(0) + a_{12}x_2(0) + a_{13}x_3(0) + b_1u(0)]\frac{\partial f}{\partial x_1} + x_1(0)\frac{\partial f}{\partial x_2} + [a_{32}x_2(0) + a_{33}x_3(0)]\frac{\partial f}{\partial x_3} + x_3(0)\frac{\partial f}{\partial x_4} \Big\} \tag{9–80}$$

The expression for the optimal control signal $u_0(0)$ is obtained by taking the partial derivative of the quantity within the braces with respect to $u(0)$. This expression is

$$u_0(0) = -\frac{b_1}{2}\frac{\partial f}{\partial x_1} \tag{9–81}$$

Substituting Eq. (9–81) into Eq. (9–80) we obtain

$$\frac{\partial f}{\partial T} = q_1(0)[x_{1d}(0) - x_1(0)]^2 + q_2(0)[x_{2d}(0) - x_2(0)]^2 + q_3(0)[x_{3d}(0) - x_3(0)]^2 + q_4(0)[x_{4d}(0) - x_4(0)]^2 + \frac{b_1^2}{4}\left(\frac{\partial f}{\partial x_1}\right)^2 + \left[a_{11}x_1(0) + a_{12}x_2(0) + a_{13}x_3(0) - \frac{b_1^2}{2}\frac{\partial f}{\partial x_1}\right]\frac{\partial f}{\partial x_1} + x_1(0)\frac{\partial f}{\partial x_2} + [a_{32}x_2(0) + a_{33}x_3(0)]\frac{\partial f}{\partial x_3} + x_3(0)\frac{\partial f}{\partial x_4} \tag{9–82}$$

If Eq. (9–82) can be solved for $f(\mathbf{c}, T)$, then the optimal control signal $u_0(0)$ can be found using Eq. (9–81).

In order to solve this problem let us assume $f(\mathbf{c}, T)$ in the following form:

$$f(\mathbf{c}, T) = k(T) - 2\sum_{i=1}^{4} k_i(T)x_i(0) + \sum_{i=1}^{4}\sum_{j=1}^{4} k_{ij}(T)x_i(0)x_j(0) \tag{9–83}$$

where $k_{ij}(T) = k_{ji}(T)$

The expression for the optimal control signal given by Eq. (9–81) can now be expressed in terms of the state variables and the k parameters using Eq. (9–83). Taking the partial derivative of $f(\mathbf{c}, T)$ with respect to x_1 and substituting into Eq. (9–81) gives

$$u_0(0) = b_1[k_1(T) - k_{11}(T)x_1(0) - k_{12}(T)x_2(0) - k_{13}(T)x_3(0) - k_{14}(T)x_4(0)]$$

If we denote t as the present time, $T - t$ as the duration of time remaining and $\mathbf{x}(t)$ as the current state, then the optimal control signal is given by

$$u_0(t) = b_1[k_1(T-t) - k_{11}(T-t)x_1(t) - k_{12}(T-t)x_2(t) - k_{13}(T-t)x_3(t) - k_{14}(T-t)x_4(t)] \tag{9-84}$$

This control law defines the configuration of the optimal control system. In carrying out the mechanization of the optimal control system the values of the time-varying gains must be obtained. These values are found by differentiating Eq. (9–83) to obtain the derivatives of $f(\mathbf{c}, T)$ appearing in Eq. (9–82); then, these derivatives are substituted into Eq. (9–82) and the terms are collected such that the resulting expression is of the following form

$$\begin{aligned} &P_0 + P_1x_1(0) + P_2x_2(0) + P_3x_3(0) + P_4x_4(0) + P_{11}x_1^2(0) \\ &+ P_{12}x_1(0)x_2(0) + P_{13}x_1(0)x_3(0) + P_{14}x_1(0)x_4(0) + P_{22}x_2^2(0) \\ &+ P_{23}x_2(0)x_3(0) + P_{24}x_2(0)x_4(0) + P_{33}x_3^2(0) + P_{34}x_3(0)x_4(0) \\ &+ P_{44}x_4^2(0) = 0 \end{aligned} \tag{9-85}$$

The P coefficients consist of terms containing the k parameters, the desired state variables, the elements of $\mathbf{A}$ and $\mathbf{B}$ which describe the aircraft, and the weighting factors $q_i(t)$.

If Eq. (9–85) is to be valid for all possible values of the state variables, each P coefficient must independently be equal to zero. Hence,

$$P_0 = P_i = P_{ij} = 0 \qquad (i, j = 1, 2, 3, 4) \tag{9-86}$$

Equation (9–86) yields the fifteen independent ordinary first-order differential equations which define the k parameters. The boundary conditions on these k parameters are

$$k(0) = k_i(0) = k_{ij}(0) = 0 \qquad (i, j = 1, 2, 3, 4)$$

Notice that the analytical design problem is thus reduced to the solution of the k equations. In general, the solution of the k equations requires the use of an analog or digital computer. Notice also that the solution of the k equations cannot be carried out until the desired state variables, aircraft parameters, and weighting factors are specified. If the desired state variables and the aircraft parameters have been specified, then, the design procedure reduces to merely the selection of a set of weighting factors $q_i(t)$ $(i = 1, 2, 3, 4)$ which will result in satisfactory system performance under the given circumstances.

The analytical design technique discussed in this problem has important advantages. In particular, multiple design considerations associated with complex dynamic situations can be included in the system design. The functional forms of the weighting factors appearing in the performance index can be chosen directly from the consideration of the time domain performance requirements. Except for

certain mathematical restrictions, the functional forms of these weighting factors can be chosen without consideration of the system stability. One disadvantage of using the present technique, however, is that, generally speaking, a unique set of weighting factors does not exist when a set of design considerations is given. In fact, the simpler the design problem from dynamic considerations, the larger the number of sets of acceptable weighting factors.

PROBLEMS

Problem B-9-1 A control system is defined by

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$$

where

$\mathbf{x} = n$ vector (state vector)
$\mathbf{u} = r$ vector (control vector)
$\mathbf{A} = n \times n$ matrix
$\mathbf{B} = n \times r$ matrix

Show that the control vector $\mathbf{u}$ will transfer the system state from $\mathbf{x}(0)$ at $t = 0$ to $\mathbf{x}(T)$ at $t = T$ if and only if

$$\int_0^T e^{-\mathbf{A}t}\mathbf{B}\mathbf{u}(t)\,dt = e^{-\mathbf{A}T}\mathbf{x}(T) - \mathbf{x}(0)$$

Problem B-9-2 Consider the control system shown in Fig. 9–18. Define $x_1 = x$, $x_2 = \dot{x}$. Suppose that the performance index is given by

$$J(\mathbf{c}, T) = \mathbf{x}'(T)\mathbf{x}(T) + \lambda\int_0^T u'(t)u(t)\,dt$$

where λ is a positive constant. Obtain u_0 and $f(\mathbf{c}, T) = \min_u J(\mathbf{c}, T)$ where $\mathbf{c} = \mathbf{x}(0)$. With $\lambda = 1$ and $T = 1$ find the state plane trajectory starting from $(2, -2)$ for the time interval $0 \leqslant t \leqslant 1$.

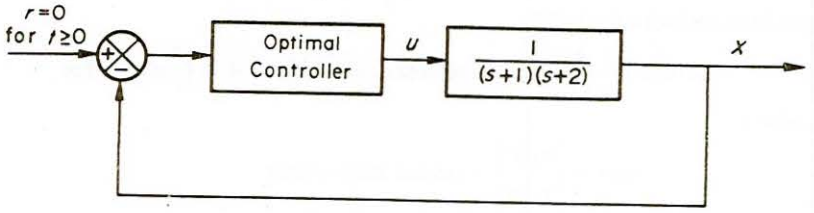


Figure 9-18 Control system.

Problem B-9-3 Obtain the Hamilton-Jacobi equation of classical mechanics from the principle of optimality. Hamilton's principle states that a particle moves so as to minimize the Lagrangian

$$\int_{t_0}^{t} L(q, \dot{q}, \tau)\,d\tau$$

where $\dot{q}(t)$ is chosen so as to transform state Q given at time t_0 into state q at time t while minimizing the Lagrangian. [L is the difference between T and V, or $L = T - V$, where T is the kinetic energy and V is the potential energy of the system.]

Problem B-9-4 Given the control system defined by

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} [u]$$

and the performance index

$$J(\mathbf{c}, T) = \mathbf{x}'(T)\mathbf{x}(T) + \lambda \int_0^T u'(t)u(t)\,dt$$

where $\mathbf{c} = \mathbf{x}(0)$ and λ is a positive constant, obtain the optimal control function u_0 and the minimized performance index, $\min_u J(\mathbf{c}, T)$.

Problem B-9-5 Consider the plant defined by

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} [u]$$

Assuming the input $\xi_1(t)$ to the control system, which has a similar configuration as that of Fig. 9–17, is a class of ramp input, and the performance index is given by

$$J(\mathbf{c}, T) = \mathbf{z}'(T)\mathbf{P}\mathbf{z}(T) + \lambda \int_0^T u'(t)u(t)\,dt$$

where

$$\mathbf{c} = \mathbf{x}(0), \qquad \mathbf{z}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \xi_1(t) \\ \xi_2(t) \end{bmatrix}, \qquad \mathbf{P} = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}$$

determine the elements of matrices $\mathbf{F}_0(T-t)$ and $\mathbf{H}_0(T-t)$ so that the performance index is minimized.

Problem B-9-6 Consider the system governed by the following equation:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} [u]$$

and the performance index

$$J(\mathbf{c}, T) = \int_0^T [\mathbf{x}_d(t) - \mathbf{x}(t)]'[\mathbf{x}_d(t) - \mathbf{x}(t)]\,dt + \lambda \int_0^T u'(t)u(t)\,dt$$

where

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \text{actual state vector}$$

$$\mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix}$$

$$\mathbf{x}_d(t) = \begin{bmatrix} x_{1d}(t) \\ x_{2d}(t) \end{bmatrix} = \text{desired state vector} = \begin{bmatrix} 1 - e^{-t} \\ e^{-t} \end{bmatrix}$$

Define

$$f(\mathbf{c}, T) = \min_u J(\mathbf{c}, T)$$

By assuming $f(\mathbf{c}, T)$ in the following form:

$$f(\mathbf{c}, T) = k(T) - 2k_1(T)c_1 - 2k_2(T)c_2 + k_{11}(T)c_1^2 + 2k_{12}(T)c_1c_2 + k_{22}(T)c_2^2$$

find the optimal control function u_0 and the minimized performance index $f(\mathbf{c}, T)$.

REFERENCES

9–1 Sibuya, Y., K. Ogata, and T. Sekiguchi, "Stochastic Optimal Control Systems," to appear in *Trans. Instrumentation and Automatic Control*, Japan.

9–2 Tsien, H. S., *Engineering Cybernetics*, New York: McGraw-Hill Book Company, Inc., 1954.

9–3 Kalman, R. E., and J. E. Bertram, "Control System Analysis and Design Via the Second Method of Lyapunov: I. Continuous Time Systems," *ASME J. Basic Engineering*, ser. **D, 82,** (1960), 371–93.

9–4 Bellman, R., *Dynamic Programming*, Princeton, N. J.: Princeton University Press, 1957.

9–5 Bellman, R. E., and S. E. Dreyfus, *Applied Dynamic Programming*, Princeton, N.J.: Princeton University Press, 1962.

9–6 Kalman, R. E., and R. W. Koepcke, "Optimal Synthesis of Linear Sampling Control Systems Using Generalized Performance Indexes," *ASME Trans.* **80** (1958), 1800–26.

9–7 Smith, F. T., "An Introduction to the Application of Dynamic Programming to Linear Control Systems," RAND *Corp. Memo.* RM-3526-PR, February, 1963.

9–8 Nahi, N. E., "Two Classes of Optimum Linear Systems," *JACC* Preprint (1964), 326–31.

9–9 Chen, Paul P., "An Algorithm for Stochastic Control Through Dynamic Programming Techniques," *IRE Trans. Automatic Control* (1962), 110–19.

9–10 Ergin, E. I., V. D. Norum, and T. G. Windeknecht, "Techniques for Analysis of Nonlinear Altitude Control Systems for Space Vehicles," Space Technology Laboratories, Inc., **2,** *Tech. Doc. Report* N. ASD-TDR-62-208, June, 1962.

9–11 Ellert, F. J., and C. W. Merriam, III, "Synthesis of Feedback Controls Using Optimization Theory, An Example," *IEEE Trans. Automatic Control* (1963), 89–103.

Index

Adaptive control system, 2
Adjoint matrix, 50–51
Adjoint system, 374–376, 419
  impulse response of, 419
  of scalar differential equation system, 375–376, 419
Aircraft landing system, 582–585
Aitken, A. C., 104, 173
Aizerman, M. A., 526
Analog computer simulation, 22–23, 182, 193–194, 309–310, 312–313, 346–347, 355–356
Analytical design, of minimum error controllers with polynomial inputs, 558–564
  of minimum error regulators, 547–557
  of noise-free optimal regulators, 529–542, 544–547
Approximation, of differential equation by difference equation, 339–343
  of input signals by polynomials, 558–559
  method of successive, 349
Associativity law, 60
Asymptotic stability, 439–441
  in the large, 441, 457
  uniform, 453–455, 457–459
    in the large, 453–455, 457–459
Athans, M., 369, 436

Barbashin, E. A., 508
Basis, 68
  dual, 74–76
  for dual state space, 376
  for n-dimensional state space, 376
  of subspace, 69
Basis functions, 295–296, 345–346
Basis vectors, 296, 301–302
Bellman, R., 104, 526, 587
Bertram, J. E., 105, 525, 587
Bilinear form, 79
Bounded input, 469–470
Bounded output, 469–470
Boundedness, uniform, 439
Butman, S., 369, 436

Cancellation, of matrix, 63
  in transfer function, 217–218, 391–392
  in transfer matrix, 389–391
Canonical form, Jordan's, 106–107
  reduction to, 132–141
Cauchy-Lipschitz theorem, 13–15
Cayley-Hamilton theorem, 245–247
Cesari, Lamberto, 525
Characteristic equation, 76–78, 93–94, 110
Characteristic polynominal, 76, 95
Characteristic roots, 77, 110
Chen, Paul P., 587
Classical control theory, vs. modern control theory, 27–28
Classical theory of dynamics, 5–8
Coddington, E. A., 368
Cofactor, 50
Column vector, 65
Complex field, 45–46
Complex matrix, 46, 55
Complex vector, 65
Complex vector space, 66
Computation of $(s\mathbf{I} - \mathbf{A})^{-1}$, 324–329
Computer control system, 2–3
Congruent matrices, 129
Conjugate matrix, 49
Conjugate transpose, 49
Connolly, T. W., 369

Conservative system, 6
Consistent system of equations, 400
Continuous-time system, equivalent discrete-time representation of 339–343
Contraction, 487, 519–520
Control energy, 529
Control law, 1
Control system, 1
  computer, 2–3
Control vector, 10
Controllability, 372
  complete output, 372, 394–396, 411–412
  complete state, 372, 381–394, 408–411, 431–433
    equivalent definitions of, 393
    s plane characterization of, 391
  δ state, 394
  relation between observability and, 413–416
  total, 372
Controllable subspace, 393, 413
Convergence of matrix series, 252–253, 284–285
Convex set, 373–374
Coordinate transformation, 112–115
Coordinates, generalized, 5–6
Cramer's rule, 91–92
Criteria for stability (*see* Stability)
Cunningham, W. J., 525

d'Alembert's principle, 29
Deadbeat response, 393
Definiteness, 79, 272–273
Degrees of freedom, 6, 8
Delta function, 18
δ state controllability, 394
Dertouzos, M. L., 44
Desoer, C. A., 173, 368, 436
Determinant, Vandermonde's, 89, 283
  Wronskian, 295–301
Deterministic system, 28
Diagonal Hermitian form, 263
Diagonal matrix, 47
  reduction to, 108, 115–118, 120–132
Diagonal quadratic form, 263
Diagonalization, of Hermitian form, 267–268
  of matrix, 108, 115–118, 120–132
  of quadratic form, 265–267
  of real symmetric matrix, 263–265
Diagonalizing transformation matrix, 113
Difference equations, approximation of differential equations by, 339–343
  classical method for solving, 364–365
  state space representation of, 218–222
  vector matrix, 218–222, 336–343
Difference of matrices, 59
Differential equation systems, state space representation of, 175–195
Differential equations, approximated by difference equations, 339–343, 365–366
  existence of solutions of, 12–15, 38–42
  Riccati, 321–322, 546, 549, 568
  uniqueness of solutions of, 12–15, 40, 42
Direct method of Liapunov (*see* Liapunov, second method of)
Discontinuity, 19–20
Discrete-time systems, 218–222, 336–343, 486–489
  stability of, 486–489
Discretization of continuous-time systems, 339–343, 365–366, 425–426
Distributivity law, 60
Domain of attraction, 441
Double pendulum, 33
Doublet function, unit, 19
Dreyfus, S. E., 587
Dual basis, 74–76
Dual space, 74–76
Dual vector space, 74–76
Dynamic programming, functional equations technique of, 542–551

Eigenvalue equation (*see* Characteristic equation)
Eigenvalues, 76–78, 110
  complex conjugate, 118–120
  of product of matrices, 77–78, 101–102
Eigenvectors, 78, 110–112
  normalized, 78
Elementary column operations, 82
Elementary matrix, 82–83
Elementary row operations, 82
Elementary transformations, 81–82
Ellert, F. J., 587
Enns, M., 44, 173, 368
Equilibrium state, asymptotically stable, 439–441
  isolated, 438
  unstable, 441
Equivalence, of matrices, 85–86
  of systems, 345–346
Ergen, W. K., 526
Ergin, E. I., 587
Estimation, of rapidity of transient response, 489–493
  of system time constant, 489–493
Existence of solutions of differential equations, 12–15, 37–42
exp $\mathbf{A}t$, methods for computing, 253–257, 315–321

Falb, P. L., 369, 436
Field, complex, 45–46
  number, 45–46
  real, 45–46
Figure of merit, 490
Finite dimensional vector space, 68–69
First approximation to nonlinear equation, 451
Forced dynamic systems, linear, 311–315, 468–470
Free dynamic systems, motions of, 151–155, 307–311
Frequency response, 27
Frobenius theorem, 244–245
Fuller, A. T., 44
Function, impulse response, 329–336
  of square matrix, 243–257
Functional equations technique of dynamic programming, 542–551
Fundamental matrix, 301–306, 337–338

Gabasov, R., 369
Generalized coordinates, 5–6
Generalized momenta, 34
Generalized momentum coordinates, 6
Gibson, J. E., 525
Gradient of scalar function, 474
Gram determinant, 376–381
Gram matrix, 376–381
Gravitational field, 42
Greenwood, J. R., III, 44, 173, 368

Hahn, Wolfgang, 525
Halmos, P. R., 105
Hamilton-Jacobi equation, 585
Hamiltonian, 6, 31
Hamilton's equations of motion, 6, 31–33
Hamilton's principle, 6, 29–33, 585
Hermitian form, 79, 262–263, 267–268, 272
Hermitian form (*Cont.*)
  definiteness of, 272
  diagonal, 263
  diagonalization of, 267–268
  indefiniteness of, 272
  rank of, 263
  semidefiniteness of, 272
  signature of, 272
Hermitian matrix, 55–56, 272–273
  eigenvalues of, 98
  eigenvectors of, 99
  indefinite, 272
  negative definite, 272–273
  negative semidefinite, 272–273
  positive definite, 272–273
  positive semidefinite, 272–273
  skew, 55–56
Hermitian performance index, 527–528
Ho, Y. C., 436
Hurwitz determinants, 448–450, 465
Hurwitz stability criterion, 448–449
  equivalence of Routh stability criterion to, 449–450, 497–499
  proof of, 466–467
Hydraulic system, 44
Hyper line in n-dimensional state space, 112

Ichikawa, K., 173, 242
Identity matrix, 47–48
Impulse function, unit, 18
Impulse response, of adjoint system, 419
  calculation of, 315
Impulse response function, 329–336, 416–418
  determination of, 333–336
Inconsistent system of equations, 400
Indefiniteness, 79–80, 272
Infinite series, 252–253
Ingwerson, D. R., 525
Inner product, 69–71, 379
Input description by state variables, 559, 574–576
Input set, 5
Instability, 441, 459–460
Integral curve, 12–13
Integral equation, 553
Invariant properties of matrix, 87
Invariant set, 458–459
Inverse matrix, 51–54, 251
Inverse of $(s\mathbf{I} - \mathbf{A})$, 324–329
Irreducible system, 418

Jacobian matrix, 451, 471
Johnson, C. D., 242
Jordan canonical form, 106–107
  reduction to, 132–141

Kalman, R. E., 105, 371, 413, 436, 525, 587
Koepcke, R. W., 587
Krasovskii's theorem, 471–472
Kreindler, E., 436
Kronecker delta, 47

Lagrange multiplier, 491, 529, 535–536, 565
Lagrange's equations of motion, 6, 30–31
Lagrange's interpolation formula, 257–258
Lagrangian, 30–31, 585
Laplace's expansion by minors, 102
LaSalle, J. P., 525
Lefschetz, S., 368, 525
Lehnigk, S. H., 526
Levinson, N., 368
Liapunov, direct method of (*see* Liapunov, second method of)
  first method of, 450–452
  main stability theorem of, 453–455
  second method of, 452–460
    proof of Routh's and Hurwitz stability criteria by, 466–468
    stability analysis of discrete-time systems by, 486–489
    stability analysis of linear systems by, 460–467
    stability analysis of nonlinear systems by, 470–485
    system design via, 489–495, 520–522
Liapunov functions, 452, 455–457
  for discrete-time systems, 486–489
  goodness of, 490
  Krasovskii's method for generating, 471–473
  Lur'e's method for generating, 479–485
  variable gradient method for generating, 473–479, 509–515
  Zubov's method for generating, 507
Limit cycle, 504
Line integral, 474
Linear dependence, 66–67, 80–81, 296–301, 379
Linear independence, 66–67, 80–81, 296–301, 379
Linear operation, 343
Linear scalar differential equation, general solution of, 301, 343
Linear system, canonical structure of, 417
  definition of, 3
  discrete-time, state space representation of, 218–222
  forced dynamic, 468–470
  impulse response of, 315
  ramp response of, 315
  stability analysis of, 442–449, 460–468
  step response of, 315
Linear transformation, 46, 69
Linearization, 451
Lipkin, H. J., 526
Lipschitz condition, 13–15, 38, 40
Loss function, 25
Lur'e, A. I., 525
  first canonic form of, 480–482
Lur'e's method, 479–485
Lur'e's theorem, 482–484

Margolis, S. G., 526
Mathematical model, 3–4
Matheson, J. E., 44, 173, 368
Matrices, congruent, 129
  equivalence of, 85–86
  orthogonally similar, 128, 157–158
  similar, 86–88
  unitarily similar, 125
Matrix, adjoint, 50
  cancellation of, 63–64
  complex, 46, 55
  conjugate, 49
  conjugate transpose of, 49
  definition of, 46
  derivative of, 64
  diagonal, 47
  diagonalization of, 108, 115–118, 120–132
  differentiation of, 64–65
  elementary, 82–83
  elements of, 46
  fundamental, 301–306, 337–338
  Hermitian, 55–56
  identity, 47–48
  integral of, 64
  inverse, 51–54, 251
  nonsingular, 51
  normal, 57–58, 98
  norms of, 78
  null, 48

Matrix (*Cont.*)
  orthogonal, 55
  partitioned, 47, 61–63
  power series, 252–253, 284–285
  rank of, 83–85
  real, 46
  real skew symmetric, 54–55, 98–99
  Riccati equation, 321–322, 546, 549, 568
  scalar, 48
  series, 252–253, 284–285
  singular, 51
  skew Hermitian, 55–56
  skew symmetric, 54–55
  square, exponential function of, 252–257, 284–285
    functions of, 243–257
    polynomials in, 243–251
  stable, 357
  state transition, 303
  symmetric, 54–55
  trace of, 47
  transfer, 322–324
  transformation, 113–118, 135–137, 148–150, 156–157, 160–170, 187, 206–207, 210–212
  triangular, 48, 104
  unitary, 57, 97–98
  unity, 47
Merriam, C. W., III, 587
Minimal polynomial, 247–251
Minimization, of integral squared error, 493–495, 520–522
  of terminal error, 529
Minimum error controller, 558–564
Minimum error regulator, 547–557
Minor, 49–50
Mirsky, L., 104
Modern control theory, 1, 27–28
  vs. classical control theory, 27–28
Modes of oscillation, 424–425
Motion, 438
Multiplication, of matrix by matrix, 59
  of matrix by scalar, 59

n vector, 65
Nahi, N. E., 587
Narendra, K. S., 436
Negative definiteness, 79, 272–273
  Sylvester's criterion for, 279
Negative semidefiniteness, 79, 272–273
  Sylvester's criterion for, 280
Newton's equations of motion, 29
Nohel, J. A., 526
Noise-free optimal regulator problem, 529–530
Nonlinear differential equation, definition of, 4
Nonlinear systems, definition of, 4
  second order, 472–473, 476, 504, 506–507, 509–511, 513–516, 524
  third order, 476–479, 481, 485, 508–509, 511–512, 516–517
Nonsingular matrix, 51
Normal matrix, 57–58, 98
Normalized vector, 70
Norms, of matrix, 78, 379
  of vector, 71–73
Norum, V. D., 587
Null matrix, 48
Null vector, 66
Number fields, 45–46
Numerator dynamics, plant with, 19

Observability, 372–373
  complete, 372
    of linear time-invariant systems, 396–408
    of linear time-varying systems, 412–413
    s plane characterization of, 407–408
  relation between controllability and, 413–416
  total, 373, 412
Ogata, K., 173, 242, 369, 587
Optimal control law, 2
  feedback, 532–533
  open loop, 532
Optimal control systems, 24–27, 527–587
  feedback, 532–533
  open loop, 532–533
Optimal policy, 542
Optimal regulator, noise-free, 529–542, 544–547
Ordinary point, 14
Orthogonal matrix, 55
Orthogonal transformation, 71
Orthogonality of vectors, 73
Orthonormal vectors, 73–74
Output vector, 10

Parametric expansion, 581
Parks, P. C., 525
Parlett, B., 173

Partial-fraction expansion technique, 203–217
Partitioned matrix, 47, 61–63
Performance index, 2, 528–530, 542, 544, 548, 560
Persidkii, S. K., 526
Phase space, 6, 8
Plant, 1
  with numerator dynamics, 19
Poles, 322
Polynominal, characteristic, 76, 95
  minimal, 247–251
  in square matrix, 243–251
Polynomial input, 558–559
  minimum error controllers for, 558–564
Positive definiteness, 79, 272–273
  Sylvester's criterion for, 277–279
Positive semidefiniteness, 79, 272–273
  Sylvester's criterion for, 280
Potential function, 42
Principal coordinate axes, 112
Principal minor, 50
Principle of duality, 414–416
Principle of optimality, 542

Quadratic form, 79, 262–263, 265–272, 274–277
  definiteness of, 272
  diagonalization of, 265–267, 274–277
  indefiniteness of, 272
  rank of, 263
  semidefiniteness of, 272
  signature of, 272
Quadratic performance index, 25, 528

Rank, of Hermitian form, 263
  of matrix, 83–85
  of product of matrices, 85
  of quadratic form, 263
Rate gyro, 37
Real bilinear form, 79
Real field, 46
Real Hermitian form, 263
Real matrix, 46
Real skew symmetric matrix, 98–99
Real symmetric matrix, 98
  diagonalization of, 263–265
  eigenvalues of, 98
  indefinite, 272
  negative definite, 272–273
  negative semidefinite, 272–273
Real symmetric matrix (*Cont.*)
  positive definite, 272–273
  positive semidefinite, 272–273
Real vector, 65
Real vector space, 66
Realization of least order system, 417–418
Reduction of system equations, 4
Region of asymptotic stability, 440
Reid, W. T., 369
Rekasius, Z. V., 525
Representative point, 5
Response, impulse, 315, 329–336
Riccati differential equations, 321–322, 546, 549, 568
Roberge, J. K., 44
Rogers, A. E., 369
Routh's and Hurwitz stability criteria, 442–450
  proof of, 466–468
Routh's stability criterion, 443–448
  equivalence of Hurwitz stability criterion to, 449–450, 497–499
Row vector, 65
Rozenvasser, E. N., 525

Saddle point, 427–428
Sarachik, P. E., 436
Scalar, 65
Scalar differential equations, 146–150, 175–195
Scalar matrix, 48
Scalar product, 69–71
Schmidt orthonormalization, 421
Schultz, D. G., 525
Schwarz, H. R., 467
Schwarz inequality, 71–72, 566
Schwarz matrix, 466
Second method of Liapunov (*see* Liapunov, second method of)
Seifert, W. W., 368
Sekiguchi, T., 587
Self-adjoint system, 375
Semidefiniteness, 79, 272–273
Separatrices, 427–428
Series, matrix power, 252–253, 284–285
Set, convex, 373–374
  input, 5
  invariant, 458–459
$s\mathbf{I} - \mathbf{A}$, inverse of, 324–329
Sibuya, Y., 587
Signature, of Hermitian form, 272
Signature (*Cont.*)
  of quadratic form, 272
Similar matrices, 86–88
Similarity transformation, 87
Simple pendulum, 7
Simplified state space representation, 204
Simulation, analog computer, 22–23, 182, 193–194, 309–310, 312–313, 346–347, 355–356
Simultaneous difference equations, state space representation of, 218–222
Simultaneous linear equations, solution of, 399–400
Singular matrix, 51
Singular point, 14
Sivan, R., 369, 436
Skew Hermitian matrix, 55–56
  eigenvalues of, 98–99
Skew symmetric matrix, 54
  eigenvalues of, 98–99
Smith, F. T., 587
Solution, bounded, 439
  existence of, 12–15, 38–42
  uniformly bounded, 439
  uniqueness of, 12–15, 40, 42
Space, dual, 74–76
  dual vector, 74–76
  state, definition of, 4–5
Square matrix, 46
  exponential function of, 252–257
  functions of, 243–257
  polynomial in, 243–251
Stability, 438–439
  analysis, of discrete-time systems, 486–489, 519–520, 524
    by first approximation, 451
    by first method of Liapunov, 450–451
    of linear time-invariant systems, 442–450, 460–468
    of nonlinear systems, 452–460, 470–485, 504–517
    by second method of Liapunov, 452–468, 470–489
  asymptotic, 439–441
    in the large, 441
  of optimal control systems, 534–535, 539–542
  uniform, 439
Stable matrix, 357
State, definition of, 4–5
State space, definition of, 4–5
State space equation, solution of, [illegible] 155, 302–315
State space representation, of input signals, 558–559, 574–576
  of scalar differential equation system, 175–195
  of system equations, definition of, 12
State transition matrix, 303
  properties of, 302–303, 305–306
State variables, 5, 16–24
  nonuniqueness of set of, 23–24
  reconstruction of unmeasurable, 373
State vector, definition of, 5
Steeg, C. W., Jr., 368
Step response, calculation of, 315
Submatrices, 47, 61
Subspace, 69
  controllable, 393, 413
Successive approximation, method of, 349
Sum of matrices, 58
Sylvester, T. T., 80
Sylvester's criterion, for negative definiteness, 279
  for negative semidefiniteness, 280
  for positive definiteness, 277–279
  for positive semidefiniteness, 280
Sylvester's interpolation formula, 257–262, 352–354
  for computing $(s\mathbf{I} - \mathbf{A})^{-1}$, 325–327
Sylvester's law of inertia, 268–270
Symmetric matrix, 54–55
System, 1, 438
  discrete-time (*see* Discrete-time systems)
  error, 547
  forced dynamic, 311–315, 468–470
  free dynamic, 151–155, 307–311
  linear, definition of, 3
  with multiple eigenvectors, 132–141
  multiple-input multiple-output, 323
  nonlinear (*see* Nonlinear systems)
System of equations, consistent, 400
  inconsistent, 400
Systems, equivalent, 345–346
Szegö, G. P., 526

Thompson, F. T., 44, 173, 368
Three axes reaction-wheel system, 37
Time-varying gain, implementation of, 533
Trace of a matrix, 47

Trajectory, 5, 303, 438
Transfer function, cancellation in, 217
  irreducible, 417
Transfer matrix, 322–324
Transformation, elementary, 81–82
  orthogonal, 71
  similarity, 87
  unitary, 71
Transformation matrix, 113–118, 135–137, 148–150, 156–157, 160–170, 187, 206–207, 210–212
  machine computation of, 141–145
Transpose, 48–49
  conjugate, 49
Triangular matrix, 48, 104
Triplet function, unit, 19
Tsien, H. S., 587
Turnbull, H. W., 104, 173

Uniformly bounded solution, 439
Uniqueness of solutions of differential equations, 12–15, 40, 42
Unit doublet function, 19
Unit impulse function, 18
Unit triplet function, 19
Unitary matrix, 57, 97–98
Unitary transformation, 71
Unity matrix, 47

V functions (*see* Liapunov functions)
Vandermonde's determinant, 89, 283
Variable gradient method, 473–479, 509–515
Vector, 65
  column, 65
  complex, 65
  components of, 65
  control, 10
  elements of, 65
Vector (*Cont.*)
  normalized, 70
  norms of, 71–73
  null, 66
  output, 10
  real, 65
  row, 65
  state, 10
Vector matrix difference equations, 336–343
  linear homogeneous, 337–338
  linear nonhomogeneous, 338–339
  simplification of, 219–222
  solutions of, 337–339, 363–364
Vector matrix differential equations, discretization of, 339–343
  homogeneous, 302–311
  linear time-invariant, 307–315
    solutions of, 308–309, 311–312
  nonhomogeneous, 306–307, 311–315
Vector space, 66
  complex, 66
  dimension of, 68
  finite dimensional, 68–69
  n-dimensional, 68–69
  real, 66
Vogt, W. G., 526

Weighting factor, 560
Wiener's filtering problem, 418
Windeknecht, T. G., 587
Wonham, W. M., 242
Wronskian, 295–301

z transform, 425–426
Zadeh, L. A., 436
Zubov's method for generating Liapunov functions, 507